Modelling Students' Behaviour And Effect in ILE through Educational Data Mining

Manolis P. Mavrikis

Doctor of Philosophy

University of Edinburgh

2008

Abstract

The long-term objective behind the research presented in this thesis is the improvement of ILEs and particularly those components that take into account students' behaviour as well as emotions and motivation. In related research, this is often attempted based only on intuition, theoretical perspectives, or guided by results from studies in the isolation of a research lab. In this thesis, an attempt was made to inform the design of adaptation and feedback components by collecting and analysing as realistic data as possible.

Guided by the belief that qualitative data analysis results can be enriched by employing statistical and machine learning techniques, this research investigated (a) key aspects of students' behaviour and their relation to their learning and (b) how their behaviour could be employed to predict students' affective and motivational states.

The first step towards this goal was to gain an in-depth understanding of students' behaviour in ILEs when they interact on their own time and location, rather than during a controlled study where the social dynamics are different. Based on the results, components of an ILE were redesigned and two Bayesian models were machine-learned; the first predicts when students need help in answering a question and the second predicts if their interaction with the system is beneficial to their learning.

In the next step, machine learning was employed in order to derive models of students' affective and motivational states based on their interactions. This was achieved by deriving decision trees based on a dataset of students' self-reports, collected during replays of their interaction. In addition, in order to take tutors' perspective into account, two different approaches were followed. The first was to elicit tutors' inferences while they are watching replays of students' interactions. Although this was not entirely successful, the difficulties which stem particularly from the fact that the tutors were asked to diagnose a situation in which they were not involved, provide insight for future work. In the second approach, decision trees were derived from a data-set of tutors' inferences, collected during one-to-one computer-mediated tutorials.

Finally, the thesis provides a detailed discussion of the difficulties encountered, implications and recommendations for future work, together with indications of worth pursuing research.

Acknowledgements

As you set out for Ithaka
hope the voyage is a long one,
full of adventure, full of discovery.
...
Ithaka gave you the marvelous journey.
Without her you would not have set out.
Ithaka, C.P. Cavafy.

The journey of this thesis would have been stopped by Laistrygonians and Cyclops, if it was not for the several sailors who, whether they wanted it or not, descended with me into the maelstorm. If you are reading this, I would like to apologise for dragging you with me, I hope during this journey you discovered something as well. I discovered loyalty, friendship and love. And if sometimes my soul set up in front of me angry Poseidons, forgive me and thank you for helping me keep my thoughts and spirit high.

I would like to thank my supervisors Antony Maciocia and John Lee for their support and assistance throughout this thesis. I am indebted to Kaska Porayska-Pomsta without whom parts of my research would have been very different. Her collaboration, unconditional support and comments helped me organise, refine and publish the ideas presented here. Working with her and Helen Pain in the EU-project LeActiveMath taught me a lot about designing and conducting studies. Also thanks to Ryan Baker for the inspiration and advice in data mining matters. In addition, I would like to thank Helen Pain and Paul Brna who often acted as mentors, helping me to identify relevant research, brought me in contact with other researchers and commented on parts of my research. My examiners, Johanna Moore and Ben du Boulay, with their critical and constructive comments helped shaping the final version of this thesis.

For the financial support, I would like to acknowledge the School of Mathematics, the School of Informatics, the University's Alumni fund and those people who supported my case for funding throughout the years. Also, Design Science for the license and in particular Robert Miner for the help in integrating WebEQ in the software that supported the data collection for the studies conducted in this thesis.

Finally, my personal thanks to my mother and my sister for their love and understanding. Despite the distance they are always in my heart, waiting in Ithaki.-

Declaration

I declare that this thesis was composed by myself, that the work contained herein is my own except where explicitly stated otherwise in the text, and that this work has not been submitted for any other degree or professional qualification except as specified.

(*Manolis P. Mavrikis*)

Σ' αυτούς που κατάλαβαν τι σημαίνουν οι Ιθάκες.

To those who perceived what the Ithakas mean.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

The integration of Interactive Learning Environments (ILEs) to support teaching and learning in many educational settings is gradually becoming a reality (e.g., Dillon and Gabbard, 1998; Anderson et al., 1995; Forbus and Feltovich, 2001). Such environments, in addition to static content, offer interactive examples and exercises. Moreover, depending on the level of Artificial Intelligence (AI) built into them, they usually adapt their material based on students' preferences or other characteristics and provide some level of feedback on students' progress, hints on their misconceptions, the solution of the exercise and suggestions of further material to study.

However, students are not necessarily accustomed to the use of ILEs. The interaction with this new medium requires new skills and ways of studying. As they often work with the ILEs in their own time and location, designers are not always aware of the exact ways they interact with the system. This results in students' using the features of ILEs in ways other than for which they were designed, despite the time spent in developing them. This is not necessarily beneficial to their learning.

The nature of the phenomenon is quite complex and the issues related to it span across different fields from the Human Computer Interaction (HCI) field to that of Educational and Motivational Psychology. The way the system is designed plays an important role, and so does the way it is introduced in the educational situations. In addition, the students' learning styles as well as their self-reflective and metacognitive skills determine their behaviour. Aspects of students' behaviour are also associated with their affect and motivation. In relation to the latter, it is well established by now, in the field AI in Education (AIEd), that adaptivity and feedback provision should take into account, apart from cognitive, affective and motivational characteristics (Lepper and Chabay, 1988; du Boulay and Luckin, 2001; Kort et al., 2001; Schank and Neaman, 2001; Burleson and Picard, 2004). In particular, AIEd researchers are interested in engaging educational systems and learners in an 'affective loop' (Conati et al., 2005). This involves (a) the detection of a learner's emotions, (b) selection of tutor actions that are beneficial to their learning and (c) the synthesis of emotional expressions which would engage the learner in a more natural interaction (Conati, 2002).

The broader goal of this thesis is to contribute towards the improvement of ILEs by understanding better the ways in which students interact with them and by modelling their emotions and motivation during the interaction. In other research this is often attempted based on intuition, theoretical perspectives or guided by results from studies in the isolation of a research lab. In this thesis, based on the principle of ecological validity, which requires methods, materials and settings of a study to approximate the real-life situation (Brewer, 2000), an explicit attempt was made to inform the design of components of ILEs by collecting and analysing realistic data.

With this overall goal in mind, and guided by the belief that qualitative data analysis can be enriched by employing statistical, data mining and machine learning techniques, the research presented in this thesis focused on

- gaining a better understanding of the students' behaviour in ILEs,
- and investigating how their behaviour could be employed to predict affective and motivational states.

While the issue of reacting to students' affect and motivation is well researched both in the field of education (e.g., Snow et al., 1996; Ames, 1992; Keller, 1983) and in the field of ILEs (e.g., Malone, 1981; Lepper et al., 1993; del Soldato and du Boulay, 1995; Rebolledo et al., 2006), little attention has been paid to the potential that the actual students' actions have to ease the diagnosis of such characteristics. While some researchers (e.g., Picard and Scheirer, 2001; Kapoor and Picard, 2005; Messom et al., 2005; Litman, 2006; D'Mello and Graesser, 2007) have investigated the issue of diagnosis based on facial expressions, voice, or other bodily measures, such methods tend to be very expensive, obtrusive and could interfere with the learning process. Since observable actions constitute the way students interact directly with the system, they are inexpensive to collect and do not require student involvement, hence limiting the feeling of being monitored. Finally, although some researchers have attempted to look into the issue of using observable student actions as a means of recognising students' affect, this is attempted either based only on theoretical perspectives (e.g., Jaques and Viccari, 2007) or following a traditional knowledge engineering approach (e.g., de Vicente, 2003). The current research investigated the potential of machine learning techniques to facilitate the diagnosis of students' affect.

Towards these goals an ILE, called WALLIS, was employed as a research tool. WALLIS is integrated into the teaching and learning of the School of Mathematics of the University of Edinburgh. By developing a mechanism that allows remote recording of the actual usage of the system, empirical data were collected from authentic student-system interactions. The data were analysed in order to identify key aspects of students' behaviour and how these relate to learning. This analysis dictated a redesign of the system before investigating aspects that relate to affect.

Having redesigned the system by improving some of the HCI aspects as well as the feedback it provides, the thesis discusses how machine learning was employed to build models of students' emotional and motivational states based on their interactions with the system. The structure of the thesis consists of the following chapters:

Chapter 2 presents the methodology and process behind this research as well as relevant theories behind the different parts discussed throughout the thesis.

Chapter 3 briefly presents WALLIS, an ILE developed partially for the purposes of this research. In addition, it briefly discusses Educational Data Mining (EDM) approaches and other mathematical techniques used later in the thesis.

Chapter 4 constitutes a core part of the thesis as it presents a detailed analysis of students' behaviour in WALLIS. Key aspects of their interaction and usage patterns are identified and relationships between measures of the student-system interaction and learning are determined. In addition, implications are drawn that play an important role in the chapters to follow.

Chapter 5 presents a redesign of the system. Based on the results from Chapter 4, HCI aspects of the system are improved and two machine-learned models are developed; one predicts when students seek help unnecessarily, and the other if their interaction overall is beneficial. The models are employed to adapt the feedback provided to the students and other interventions. In addition, the redesign, facilitates the investigation of affective states in the next chapter.

Chapter 6 constitutes the other core part of this thesis. It investigates the use of machine learning in predicting students' emotional and motivational states. First, the investigation focuses on a dataset of students' self-reports collected during replays of their interaction. Whilst the methodology seems promising and the results satisfactory, the fact that the data are based on student's perspective introduces bias in the model. In order to take tutors' perspective into account, two different approaches are presented. The first, attempts to elicit data by asking tutors to infer students' affective states during replays of their interaction. This met with difficulties which are discussed in detail together with insights for further work. In the second approach, tutors' inferences collected during one-to-one computer-mediated tutorials are analysed. Despite the highlighted difficulties the machine learning methodology seems promising again.

Chapter 7 draws some final conclusions in the light of all the issues explored in the previous chapters. It also presents the contributions of this thesis and highlights the limitations and challenges faced. Finally, it discusses possible further research that can alleviate these limitations as well as further issues that emerge from the work described here and are worth pursuing.

A detailed chronology of the studies and data collection that guided this thesis overall is included in Appendix A. The following papers have been published in connection with the research presented in this thesis. Parts of Chapter 3 and particularly the logging mechanism of WALLIS developed to facilitate the data collection are described in (Mavrikis, 2005). The parts of Chapter 4 and 5 that describe the redesign of the system and some of the observed interactions appear in Mavrikis and Maciocia (2003b). A short version of Section 6.2, which is concerned with the investigation of students' self-reports, was presented in Mavrikis et al. (2007). Finally, parts of Section 6.3 are expanded from the author's contribution to Porayska-Pomsta et al. (2008).

Chapter 2

Related Research and Methodology

2.1 Introduction

This chapter constitutes the background of this thesis presenting the related research and discussing other aspects that are considered necessary prerequisites for the reader to understand and appreciate parts of the research described here. First of all, Section 2.2 presents the general methodology behind the research. After having established the methodology the chapter focuses on related research that inspired the investigations in this thesis. As mentioned in the Introduction the issues investigated relate to aspects of human emotions, motivation and behaviours and, as will be explained better in the following sections, the context plays an important role when investigating such complex issues. Therefore, before dwelling on the details of the related research, it is worth describing aspects of the context in which this research is situated.

First of all, the term Interactive Learning Environments (ILEs) refers to educational systems where learners, apart from studying static content, can interact with examples and exercises. Whether these systems are available over the internet (and hence often generally referred to as *eLearning*) or locally to classrooms' or students' personal computers, there is an implicit assumption behind the thesis that students interact independently with these systems rather than in classroom setups assisted by the teacher or collaboratively with other students. Another assumption is that the ILEs under discussion have a common feature; they try to engage the student in a meaningful interaction with the learning material rather than simply provide content. Usually constructivist learning theories (e.g., Duffy and Jonassen, 1992; Golden, 1990) are hidden behind this approach and therefore an effort is made to provide stimulating learning opportunities to the student.

Depending on the level of Artificial Intelligence (AI) built into the ILEs and the exact pedagogical perspectives taken, these systems usually adapt their material based on students' preferences or other characteristics and provide some level of feedback on students' progress, hints on their misconceptions, the solution of the exercise or suggestions on which further material to study. These systems are often called Intelligent Tutoring Systems (ITS) to highlight their intelligent features or Adaptive Hypermedia Systems (AHS) to highlight their adaptive features and the fact that they are offered online and combine hypermedia facilities to deliver the content.

The exact distinction between each type of system is not very clear as they can differ and be similar along several dimensions. The dimensions of particular importance for this discussion are (a) which student characteristics or behaviours these systems adapt to, (b) the help they provide and (c) the amount of system versus student-control during the interaction. Most typical ITS for example employ AI techniques to provide feedback to the learners by tracing their problem solving steps or modelling their knowledge in a particular problem solving situation. Although the student controls the interaction in terms of the feedback provided by requesting help, the exact sequence of the material (e.g., the exercises) are predetermined or decided based on students' misconceptions or other characteristics. In typical eLearning systems the list of concepts covered is longer and the learner has more control over which material to study but usually the help offered is limited. These systems are also often referred to as CBT from the fact that they offer computer-based training. AHS try to alleviate problems of eLearning systems by adapting the material presented to students' preferences and learning styles and helping the student manage the size of the available concepts by adapting the links between material, monitoring their history and suggesting further material to study.

In the last few years there has been a tendency in the field to combine features of content-based approaches from CBT with adaptive educational strategies from ITS

(Brooks et al., 2006; ADL, 2001) building what is often referred to as advanced e-learning systems. This way, while learning can be conceived as an individual, mostly active and self-regulated, situational process initiated by learners themselves (Duffy and Jonassen, 1992), the importance of the system's ability to provide help when the student is not able to progress is also understood.

In this thesis, the term ILE is used to refer to systems that combine features from all the above areas. Although a particular one (described in the next chapter) will be employed as a research tool, the aspects of the interaction that will be investigated are similar to many state-of-the-art environments whether web-based or not.

Finally, another contextual factor is worth mentioning here. The investigations are focused on higher education students studying mathematics. Both the subject matter, and the fact that all the studies in this thesis are performed with slightly older students, shape many of the investigations. For example, studying mathematics has been associated with higher emotional complexity and feelings of anxiety (known as 'math anxiety' Baylor et al. (2004); Op't Eynde et al. (2001)). The age of the students has been shown to determine their metacognitive skills. Older students tend to be better (Schoenfeld, 1987) and they expect to have more control over the educational situation (Knowles, 1975). Contextual factors such as the above have to be taken into account while drawing comparisons with other studies.

After presenting the general methodology in Section 2.2 the related research is divided as follows. Section 2.3 outlines research covering theories of emotions, motivation and their relation to learning. While reviewing each of them is beyond the scope of this thesis, some of the basic ideas behind the theories that play an important role later in this thesis are discussed. Afterwards, based on the fact that the thesis is concerned with students' behaviours one of the prominent aspects of student actions, help-seeking, is discussed in detail. Help-seeking is one of the behaviours that is found both in the context of classroom education and in ILE. In addition it has been investigated in detail (in the classroom context) and the relevant research plays an important role in the rest of this thesis. Finally, Section 2.5 revisits the issue of emotions and motivation in the context of ILE in particular and outlines the research approaches of other research, highlighting the relevance of this thesis in the field.

2.2 General Methodology

The thesis borrows methodology ideas and principles from different fields. The underlying methodology in terms of how the research is conducted overall is inspired by Persistent Collaboration Methodology (PCM) (Conlon and Pain, 1996) in the field of applied AIEd and by socio-technical (Sharples et al., 1999), user-centred (Krutchen, 1999) and other approaches that advocate iterative development (Larman and Basili, 2004). In particular, PCM provides a suitable framework for the research of this thesis because as Conlon and Pain (1996) highlight:

> "PCM is inspired by action research, which (despite the danger of producing results that are difficult to generalise) is the only transformational research method in education that looks at changes in the environment, addresses practical questions, and is appropriate when a new approach is to be integrated into an existing context." (Conlon and Pain, 1996)

The methodology advocates that when a new approach (for example the use of ILE) is to be integrated into an existing context (for example the educational situation) it is essential to do so incrementally and through phases of an '*observation*, *reflection*, *designing* and *action*' cycle (see figure 2.1).

The *observation* part of the cycle fits particularly well with the nature of the phenomena investigated in this thesis. Typically, socio-emotional contexts dictate the use of exploratory observational approaches (Porayska-Pomsta and Pain, 2004). In addition, as will become clear in the next section, the fact that learning, emotions, motivation and students' behaviour are all intertwined makes it hard to isolate issues behind each one of them. By following an iterative methodology it is easier to at least control or be aware of some of the factors that can influence a study.

The system employed behind the studies of this thesis, WALLIS (described in the next Chapter), is also built following an iterative design methodology and fits particularly well the research principles here. By iteratively designing the system and feeding results from research such as the current one to its development, it was established that the system is overall effective in terms of students' learning. Having established this, investigating aspects related to interaction with the system, learning, affect and motivation became more plausible.

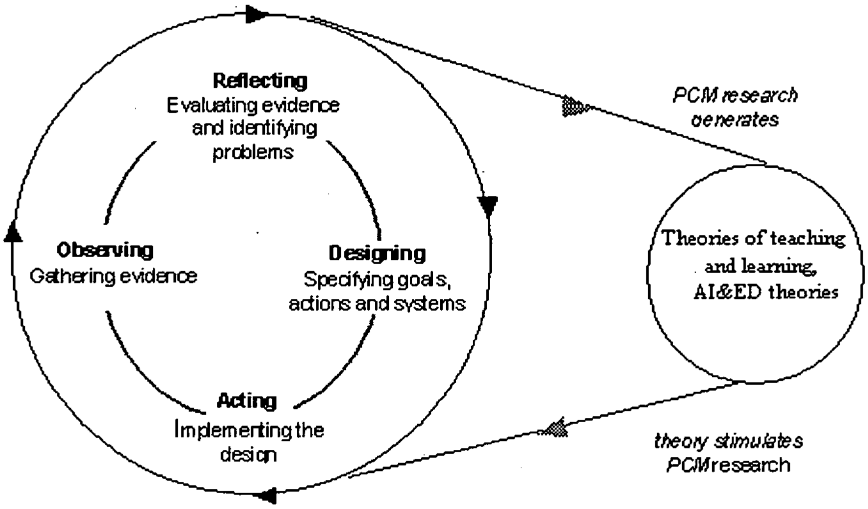


Figure 2.1: The cycles of PCM (from Conlon and Pain, 1996).

The effectiveness of the methodology becomes more apparent when one considers that results from educational research do not always generalise to the field of ILEs (Martinez-Miron et al., 2003; Aleven et al., 2003). Especially when dealing with issues that pertain to affect, emotions and motivation, theories and observations derived in the classroom are not necessarily applicable in this new medium, where students work on their own time and location. By integrating an ILE and conducting *observations* of its actual use, the issues under consideration can be observed in their actual context. While the importance of conducting such empirical observations is generally understood in the field (e.g., Conati and Maclaren, 2004; Porayska-Pomsta and Pain, 2004; Luckin et al., 2006) not many studies investigate the actual use of ILEs. Such studies appear more often in the HCI field (e.g., Peters, 1998) to determine the ways users interact with the systems under realistic situations. This is the reason that the thesis establishes first a deeper understanding of the ways students interact with the system, particularly when they are working on their own time and location rather than in a research lab.

The analysis above constitutes the *reflection* phase while the problems identified initiate the next phase of PCM, the *design*. In general, Conlon and Pain (1996) expect this process to contribute to the 'wheel' of knowledge (on the right of Figure 2.1) enabling the study of theories of teaching and learning as well as AI techniques and tools. In the case of the current research, it enables the investigation of how machine learning can be employed to build models of affective and motivational characteristics based on students' actions. PCM also expects that the development of theories or models will stimulate further cycles of PCM. It is obvious that this cycle cannot end within the constraints of a thesis like the current one. However, the research findings can influence the redesign of the system setting in motion further cycles of observation, reflection, designing and action.

All the above describe the overall methodology and inspiration behind this thesis. The methodology behind each specific part of the thesis is described in the corresponding chapters. Here, it is worth mentioning that the methodology (and goals) of this thesis are shaped by the following two beliefs. First of all, as already discussed above, the nature of the phenomena that the thesis deals with dictate the use of exploratory, observational approaches. The main principle however behind any approach is take into consideration their 'ecological validity'. For example, as mentioned in the Introduction, the environment, materials and settings should approximate as closely as possible the the real-life situation under investigation (Brewer, 2000). Despite the difficulties involved in such a methodology, it helps eliciting more realistic data, the analysis of which can lead to more solid results. The second belief is that machine learning and data mining techniques can be used to enrich results from qualitative research. The influence of both of these beliefs will become apparent.

2.3 Emotions, Motivation and Learning

It is well known that emotions often interfere with mental life. This is illustrated in Damasio's seminal work where it is argued that emotions and feelings play a crucial role during reasoning and knowledge acquisition in general (Damasio, 1994, 1999). As early as 1913, Dewey (1913) highlighted the importance of students' motivations in learning and recognised the importance of interest and effort. Similarly, according to many researchers (e.g., Snow et al., 1996) cognition, motivation and emotion are intertwined components of learning.

In addition, research has established that human tutors pay at least as much attention to the achievement of *motivational goals* as to cognitive ones during teaching (Lepper et al., 1993), and they try to respond or adapt their teaching according to students' emotions and motivation. In the field of motivational psychology, Weiner (1992) points out that humans "continuously consider 'why questions' and if possible come up with answers" in order to explain the behaviour of others but also to adapt their own behaviour. This is related to *attribution theory* which is concerned with perception of causality or the perceived reasons for a particular event's occurrence. Weiner proposes that motivational phenomena could be analysed by examining people's causal attributions for their successes and failures and highlights that, not only do clear individual differences exist among people in their tendencies to make certain attributions, but also attributions can be situationally induced (Weiner, 1990). In the context of education, this implies that both tutors and students tend to look (whether implicitly or explicitly) for cause-effect relationships affecting their behaviour during a learning situation. The relevance of the above appears particularly when one considers helping and help-seeking where the perceived cause of a student asking for help can determine the tutors' behaviour but also, the way a student perceives the cause of their need for help (e.g., personal failure) determines whether to seek help or not. This will become more apparent in Section 2.4 which discusses help-seeking in particular.

Weiner presents in detail determinants of the evaluation of individuals in achievement contexts. One intuitive, but often neglected, characteristic is *effort*. Weiner (1992) cites literature (pp. 334-336) that clearly indicates that apart from high achievement, the other determinant of positive teacher evaluation is effort. These examples show that when marking exams, high effort or motivation is rewarded more for success and punished less for failure. Lack of effort, accompanied by high ability, elicits the greater punishment (in this case the student is clearly responsible for failure). Although the effects of such an appraisal method are not discussed in detail in the literature, the fact that it is natural in (good) human teachers indicates that it could also be effective.

This assumption is strongly supported by results in a rather different research field, that of *belief systems*. Dweck (1999), while summarising the research on belief systems and their relationship to motivation and achievement, points out that there are two views of intelligence. People with an entity view, who consider intelligence to be fixed, and people with an incremental view, who believe intelligence or ability to be changeable. Students with an entity view of intelligence tend to develop goals that are based on performing better than others and on avoiding failure since they believe that circumstances are beyond their control and give up easily. They often avoid challenging activities or they attempt extremely difficult things so that they have an excuse for failure. Following failure, they may switch to an easier task or stop trying altogether. On the other hand, students with an incremental view cultivate their intelligence through effort, task involvement, and strategy development and tend to develop mastery goals with respect to achievement. These students are interested in learning and mastering challenges. Following failure, they remain confident that they can succeed by revising their strategies and increasing their efforts and believe that effort, through increased learning and strategy development, will actually increase their intelligence. Dweck's findings suggest that teachers should encourage students "to relish challenge and effort, and to use errors as routes to mastery" (Dweck, 1999, pp4) and emphasises the importance of praising effort. When teachers praise effort and strategic behaviours, students develop learning goals and a mastery orientation instead of performance goals that may lead to learnt helplessness.

Similar implications can be derived from the the large body of research that concentrates on students' *epistemological beliefs*, *learning styles* as well as *achievement goals*. Epistemological beliefs are the students' belief about the nature of knowledge in general. Regardless of the exact dimensions that the models behind epistemological of beliefs have (Schommer, 1993; Kardash and Scholes, 1996; Schoenfeld, 1983) it is understood that these beliefs may operate as a control that determines students' behaviour. Similar is the effect of different learning styles and achievement goals.Achievement goals can be broadly separated to *learning goals* and *performance goals* Newman (1998). Learning goals guide a students' actions making them seek challenge and mastery of tasks while performance goals stress the demonstration of high ability and avoidance of judgements of low ability.

While the above demonstrate the importance and complex effect of emotions and motivations in learning, the accurate explanation of behaviours is often a difficult task.

Goleman (1996) presents examples of research that demonstrate that although the ability to empathise with other people's emotions is taken for granted, people differ a lot. Weiner (1992) points out that the naive observer often answers a 'why question' with a trait, a stable characteristic, of the acting person, neglecting that there may be many other determinants of an action such as states, moods, emotions, conscious thoughts, unconscious attitudes and the environment itself. Theferore, students' behaviours should not be interpreted out of context, but have to be investigated in the classroom context, and the nature of the tasks. This is discussed in more detail in Section 2.4.3.3.

In addition, the exact approach to motivate students is not well defined let alone how to detect their emotions and motivation. Although the theories provide recommendations for teachers and researchers in the field they are often too vague, too contradictory, and too abstract to be really useful (particularly from a computational point of view).

Research suggestes that even if an effective approach in the classroom is followed to motivate students, their perceptions of the classroom and their individual motivational orientations and beliefs about learning will play an important role in their cognitive engagement and performance (e.g., Ames and Archer, 1988). This is particularly true for adult learners in general and in particular learners working in an ILE where they are supposed to have more control over their learning.

While the transfer of control to students has the added benefit of achieving higher levels of learner satisfaction (Knowles, 1975) this has to be achieved carefully. Particularly in the context of adult learners (Knowles, 1975) indicates that learners need to know why they need to learn something and they need to be responsible for the decisions they make. In addition, the environment should foster self-directed learning and treat the students as capable of self-direction (Knowles et al., 1998). In these cases, the metacognitive strategies for planning, monitoring, and modifying students' learning strategies as well as the control of the effort they put into tasks, and the actual cognitive strategies that they employ in order to learn, remember, and understand the material are very important (Zimmerman and Pons, 1988). All these aspects are related to the construct of *self-regulated learning* which plays a particular role in learning with

ILEs due to the fact that the learner usually plays a more active role.

Relevant studies provide evidence that a major cause of under-achievement is the inability of students to control themselves (Krouse and Krouse, 1981; Borkowski and Thorpe, 1994; Zimmerman, 1994). In today's world and particularly in the context of higher education, where students continually have to acquire new knowledge and assess its usefulness, self-regulation is one of the skills students have to acquire in order to be good (continuous) learners (Bransford et al., 1999).

In social environments, a manifestation of self-regulated learning is the ability of the student to seek assistance by asking questions (Newman, 1994; Aleven et al., 2003; Karabenick, 2003). Therefore, one of the most prominent aspects of students behaviours in any educational context is help-seeking. Many researchers provide evidence that help seeking is one of the most complex processes compared to other strategies of self-regulated learning particularly because it is a social strategy. The next section discusses this issue in more detail.

2.4 Help-seeking and feedback provision

Seeking help when needed, reflecting on, and interpreting the given help are important skills in their own right as they are the means by which one can acquire further knowledge to that given. They are also the means to learn what is not clearly understood.

In the context of this thesis, help-seeking plays a predominant role as it is one of the behaviours in ILEs that has not been investigated in detail and therefore the exact way students seek help is not very well known. Studies in the educational field provide evidence that many students either do not seek help effectively, or avoid seeking help (Nelson-Le Gall, 1987; Karabenick, 1988). Others show that students with high prior knowledge are more effective help seekers. The worrying aspect is that students with low prior knowledge, who need help most, are not receiving it and face more and more difficulties while learning. Particularly in classroom settings there are various reasons for not seeking help: social constraints (Nelson-Le Gall, 1981), personal beliefs (Dweck, 1999) and expectancy of the outcome (perceived causes of outcomes) (Weiner, 1992; Pintrich, 2002) as well as a more general orientation towards learning or performance (Arbreton, 1998). Which of these also play a role in ILEs is not very clear, but certainly some of the aspects seem related. This also implies that perhaps students' help-seeking behaviour has the potential to contribute in the modelling of affective characteristics as well as to drive further adaptation and intervention.

The following subsections discuss in more detail some of the related research from classrooms, the relation of help-seeking to learning in ILEs, and the factors that seem to influence help-seeking.

2.4.1 Types and models of help-seeking

An attempt to explain the individual differences in students' help-seeking behaviour is the work by Nelson-LeGall (1985). Nelson-Le Gall (1981) developed a model for traditional educational settings. First she explains that help seeking is useful under certain conditions and that students are not necessarily immature or dependent when they seek help. Similarly, students are not necessarily mature or autonomous when they do not seek assistance from others. Nelson-LeGall (1985) then differentiates between *executive* and *instrumental* help-seeking. Executive help-seeking supports performance orientation goals and aims at completing the task by requesting the tutor to answer the question. In contrast, instrumental help-seeking involves requesting help that aims to demonstrate or explain the method by which the problem can be solved, allowing the student to retain responsibility for the solution and to acquire new knowledge. This way the help seeker not only can remedy their immediate problem, but also ensure long-term autonomy. Nelson-Le Gall (1981) suggested that students who attribute failure to internal factors such as lack of effort are more likely to exhibit this type of help-seeking.

Expanding the above Newman (1994) based on Nelson-Le Gall (1981) defines a general model of help-seeking where the student:

1. becomes aware of task difficulty (or need of help)
2. considers all available information (task demands, costs and benefits etc.) and decides:

   - the necessity of the request

- the content or form of the request

- the target of the request

3. expresses the request in a suitable way according to the circumstances and

4. processes the help that is received in such a way that the probability of success in subsequent help-seeking attempts is optimised

The model may seem simple at first sight but a number of the actions and decisions involved are cognitively and metacognitevely demanding and involve lots of aspects for a student to take into account which are oversimplified. For example, being aware of the task difficulty is a metacognitive task. Students have been found to either overestimate or underestimate task difficulty, let alone the need for help. As Newman (1998) puts it, determining the necessity of asking for help rather than choosing an alternative strategy (e.g., retrying the question) is a function of students' reflecting on their sense of task difficulty in relation to their knowledge, their beliefs and feelings about themselves (from example their confidence) as well as achievement goals. When it comes to evaluating the cost and the benefit of a request, several aspects come into play. First of all, prior knowledge plays an important role. For example, students with greater knowledge of the domain are more capable of asking more related questions but also understand the difficulty of the task. Since this requires metacognitive skills from the students, age becomes another factor. Schoenfeld (1987) indicates that children are not very good in metacognition but they get better as they get older. However, even between same ability and age students other affective characteristics as well as personal or even epistemological beliefs and other situational and social factors (e.g., who else is present in the room) determine students' help-seeking behaviour. A relevant construct here comes from the area of social linguistics; *face* (Brown and Levinson, 1987). Face refers to a person's need to maintain autonomy and to be approved by others. Brown and Levinson (1987) also highlight the particular importance of the cultural and situational context in the linguistic realisation (in this case the feedback provided during a tutorial dialogue). Finally, it is worth being aware that 'goals that students bring from home as well as goals that are infused in the classroom significantly influence the academic help-seeking process' (Newman, 1998).

2.4.2 Help-seeking in ILE

The aforementioned offer a brief review of the literature of help-seeking with the perspective of classroom or one-to-one tutorials. It was made clear that help-seeking is a complex process and cannot be interpreted only as a simple cognitive process. As Newman (1998) describes, if help seeking could be viewed solely from a cognitive perspective it would be less complicated but "raising one's hand, admitting difficulty or failure, and asking for help are 'social-interactional' exchanges". As such, not all the aspects mentioned above will apply in the context of ILEs, and the ones which apply may take different forms and play different roles. For example, it was established in the previous section that quite a few affective factors play a role but not all of these have the same impact when thought of in the context of ILEs (for example the presence of other students in the classroom).

In addition, the previous section outlines the importance of help-seeking in learning but also in determining how a tutoring session will unfold. In the context of ILEs, Wood (2001) and Aleven et al. (2003) indicate that help-seeking is an even more important aspect of the interaction. It helps in overcoming the problems that arise from the fact that these environments do not have as much variety of bandwidth (VanLehn, 1988) as face-to-face interactions in order to assess if students are on task, confused, thinking or in the process of answering.

A common solution in ITS is to leave the decision about when to seek help to the student (Wood, 2001; Aleven and Koedinger, 2000). The (system) tutor then decides what help to provide. Other systems provide feedback when students attempt to answer a problem or steps of it. Whatever the exact approach, there is an implicit assumption again that places more responsibility for regulating the tutorial interaction on the student. This is usually justified based on the fact that it allows learners to be more actively involved in the learning situation and creates opportunities to develop skills of how to regulate their own learning (Wood, 2001), something that from a constructivist point of view is desirable. On the other hand, this difference of the student-driven interaction imposes a danger. Not only is there an additional cognitive demand on the student who has to seek help (Wood, 2001), but also many aspects of the student's affective characteristics come into play leading to help abuse or avoidance (Aleven and Koedinger, 2000). These could potentially hinder learning. The issue has been also highlighted by Baker (2005) who discusses help-abuse by students and coined the term 'gaming the system' to describe student behaviour that aims to complete problems and advance through the material by systematically taking advantage of properties and regularities in the system to complete a task, rather than by thinking through the material.

Despite the substantial effort researchers and developers put into developing help facilities little is known (up to recently) about the actual use of students' help-seeking behaviour in ILEs and its relation to learning. The few related studies that look particularly at this issue report conflicting results. On one hand, (Wood and Wood, 1999) found a positive correlation between help seeking and learning and mention only mild instances of help abuse. Aleven and Koedinger (2001) however, find a negative correlation between help seeking and learning gain when partialing out pre-tests, implying that any learning advantages due to the more frequent use of help messages were not sufficient to enable the more frequent help users to overcome these difficulties. Renkl (2002) compared two versions of an ILE for studying examples, one that offered help and one without, and found that all learners except those with high prior knowledge had higher learning gains if if they sought help more often. This result is consistent with Wood.

Given the paucity of research in realistic situations (i.e., with actual ILEs as integrated in the educational situation) where there are indications that students behave differently, extrapolating results from different contexts can lead to erroneous conclusions (Aleven et al., 2003). Therefore more empirical research is needed in the context of ILEs. Several characteristics of students' behaviours are influenced by their awareness of being monitored, commonly referred to as Hawthorne effect (Gillespie, 1991). This is particularly true in the case of the help-seeking behaviour which is a highly social behaviour. Aleven et al. (2003) highlight that a challenge for researchers should be the identification of factors that influence help seeking behaviour and the effectiveness of help, as well as the identification of how a given help system is used in different contexts. The next section examines some of the factors that influence help-seeking.

2.4.3 Factors Influencing Help-seeking

In general, a huge array of factors influence the decision to help or not to help another person, as well as to seek or not to seek help. Reverting to motivational psychology, Weiner (1992) finds that among the determinants of help giving are the perceived benefit of the recipient of the aid as well as the values and norms of the culture and the environment[1]. But most important, in the context of classroom behaviour, it has also been reported (see Weiner pp.314-317 for references) that the perceived cause of the need for help is the most important determinant on the decision about whether or not to help a student who requests help. Weiner provides many related examples concluding that causes perceived as subject to students' personal control give rise to neglect (or are treated with 'punishment'). However, causes perceived as uncontrollable tend to generate help. "Hence, there is an association between a dimension of causality (controllability) and a behavioural consequence (help versus neglect)." (pp 317).

In terms of the actual request for help, Aleven et al. (2003) offers a comprehensive review of factors which influence help seeking. Here an overview of some of them and some additional ones inspired from Weiner (1992); Bartholome et al. (2004) and others are discussed. In Section 4.3 an attempt is made to investigate some of these empirically.

2.4.3.1 Prior knowledge

It is well known that prior knowledge is often a good predictor of learning and performance in many fields (e.g. Bloom, 1976; Dochy, 1996; Tobias, 1994). It is also a factor that can be easily identified and diagnosed. However, in terms of determining help-seeking the results are not so clear. Even studies on help-seeking in traditional classroom settings provide contradictory results. For example, Puustinen (1998) confirms the fact that learners with low prior knowledge are least effective in asking for help. Scardamalia (1992) on the other hand, shows that students with less prior knowledge ask more appropriate questions and learners with high prior knowledge may over-

[1]There are other factors that are not so related to tutoring. For instance, aspects that relate to the motive of the help giver can be ignored in this context under the assumption that teachers (or even computer tutors) will provide help based on other factors than their own benefit or cost of help giving.

estimate their understanding.

In the context of ILE, Wood and Wood (1999) found significant interactions between level of prior knowledge, performance, and frequency of help-seeking in an ILE. Overall, learners with more prior knowledge asked for help less often. However, they manifested more effective help-seeking behaviour and tended to seek help after making an error much more frequently than learners with low prior knowledge. Learners with less prior knowledge made more errors but sought help less often. Unfortunately it is these students who need it most. Wood and Wood's results are rather counterintuitive. If students with low prior knowledge made more errors and self-corrected less, then the question of why they avoided asking help remains. Aleven and Koedinger (2001) however, find no significant interaction with prior knowledge. It seems once again that such interactions need more detailed research and perhaps they are specific to the domain, student and system.

In terms of the differences with classroom behaviour, or one-to-one situations it is possible that tutors adapt (guide the interaction) more when tutoring low achievers than high achievers (Slavin, 1987; Chi et al., 2001). Therefore, the interaction is different when compared to ILEs where help is under student control. In addition, in terms of the differences between studies in ILEs, the results discussed in Wood and Wood (1999) are from lab studies where students are called to participate, whereas the results in Aleven and Koedinger (2001) are from situations where students work with the system in the classroom.

2.4.3.2 Affective and motivational characteristics

The discussion in Section 2.3 suggests that emotions and motivation play an important role also in help-seeking. For example, Arbreton (1998) showed an effect on help-seeking strategies. Learning-orientated students seem to ask for instrumental help much more frequently, whereas performance-orientated students tended to ask for executive help.

In Section 2.3 *effort* was identified as an important factor. Nicholls et al. (1990); Rollett (1987) provide examples of the influence of the students' tendency to avoid effort in help-seeking. Although invested effort is often associated with help seeking, that is, the more questions one is asking the more effort they are exerting (Bartholome et al., 2004; Rebolledo-Mendez, 2007), this depends on the kinds of help-seeking. For example, if a student is requesting help just to avoid errors or is trying to get the right answer just to proceed to the next task, then it is unlikely that this behaviour demonstrates that they are exerting effort. Another variable that has been shown to affect learning is *interest*. For example, Schiefele (1991) showed that high interest is associated with a deeper level understanding. Intuitively, it is also likely that interest affects students' help-seeking in ILE.

Results from educational research suggest that self beliefs as well as epistemological beliefs mentioned in Section 2.3 also play a role in the way one seeks help. Similarly one could hypothesise that students' learning styles and achievement goals could be playing a role. Newman (1998) highlights that the help-seeking literature is full of evidence of how achievement goals influence help-seeking. Briefly, it is worth mentioning that in general students with learning goals tend to seek task-related information and to confirm their work while students with performance-oriented goals tend to immediately request answers without first attempting the problem on their own or asking for final answers (Newman, 1994, 1998).

However, the extent that these findings generalise to ILEs is not very clear. Bartholome et al. (2006) provide some evidence of the effect of epistemological beliefs in learning. They demonstrate that more sophisticated beliefs in knowledge being unstructured and flexible results in a higher amount of help-seeking and suggests that the mechanism by which epistemological beliefs might influence performance is in the amount of metacognitive activity. Baker et al. (2005) however do not find any connection between students having performance goals and their help-seeking behaviour. Baker's results suggest that although performance-oriented students manifest different behaviour (e.g., they work slower and avoid errors more than others) they do not attempt to 'game the system'. As Baker et al. (2005) also point out the differences between researchers and results can often be caused by the exact definitions and tools used to identify students with performance or learning goals. Similarly, there could be differences in the exact definitions of help-seeking, the kinds of help the system provides and so on.

2.4.3.3 Contextual Goals / Task Orientation

The last subsection highlights the fact that students orientations and beliefs play an important role in their behaviour and particularly their help-seeking. As mentioned in Section 2.4.1, these beliefs are rarely independent of the overall goal that the classroom or other contextual factors (culture, society, family) impose on students. As Eccles and Wigfield (1985) also note, teacher expectations, perceptions and implicit or explicit messages affect students behaviour. Newman (1998) also underscores that students' personal goals might moderate effects of contextual goals and provides evidence that when both contextual and personal goals emphasise performance, students' reluctance to seek help seems to be reinforced. However, when performance-oriented students are placed in a context that promotes learning, the cues from the environment help them overcome their personal tendencies.

Despite the different means (classroom vs computer education) it seems that these expectations have a powerful influence on student's interactions with ILEs as well and can potentially hinder learning.

2.4.3.4 Aspects of the ILEs and students' familiriarity with it

As with the overall contextual goals, similarly the context of students interacting with a learning environment is the learning environment itself and this is bound to affect their behaviour. First of all, HCI factors can affect students' emotions and their beliefs about the system itself. For example, certain features failing to provide appropriate feedback can cause frustration to students or make them disregard the system. In addition, the exact content of the feedback also plays a role. Studies on feedback messages McKendree et al. (1998) show an effect of aspects such as the help message referring explicitly to students' goals, the interactivity (i.e., being asked something rather than just providing help) and other aspects of the feedback, even its length.

In addition, Aleven et al. (2003) speculate that although little research is available about the influence of learners familiarity with the system on their help-seeking behaviour, it is possible that users need to know about the help functions offered within an ILE before they can use them in an appropriate way. This is also suggested by relevant data from early pilots with the system used for the purposes of this thesis, where it was observed that students who become familiar with the system and its functionalities behave differently than when they first interact with the system. On the other hand, the data also suggest that while some of the students become proficient with using the system, they also develop a tendency to abuse the help that it offers and use its functionalities to avoid effort. This is also suggested in the research mentioned previously in the form of 'gaming the system' (Baker, 2005).

In the spirit of ecological validity it seems that data from empirical studies will only be more valid when students are already aware of the system's capabilities and when the system has been tested for several features such as the feedback it provides.

2.4.4 A model of good help-seeking behaviour in ILEs

While the current research was ongoing, Aleven et al. (2004) developed a model of help-seeking in ILEs that shares some general traits with models of help-seeking in the classroom (such as the Newman (1994) mentioned in the previous section). This model describes the ideal help-seeking behaviour and attempts to promote instrumental help seeking. The idea in Aleven et al. (2004) is that deviations from this model could be treated similarly to buggy rules for cognitive processes by providing appropriate feedback about their actual help-seeking process. The feedback is supposed, not only to help students interact more efficiently with the ILE (and therefore learn more) but also to learn how to be good help-seekers. Although the model is rather specific to the CMU tutors it has some interesting general parts that could be applicable in ILEs in general.

According to the model (see Figure 2.4.4) in a given step a student has three choices: (a) to attempt to answer (b) to go to the Glossary (in order to get de-contextual help) or (c) to request a hint. If after spending some time and if the step looks familiar, then students who have a good idea of how to solve the problem should attempt to answer. On the other hand, if the step looks familiar but they do not know what to do, they should look at the glossary to explore definitions and formulae that may be helpful. If after using the Glossary they still have doubts, then they could ask for help from the tutor. If from the beginning the step is not recognisable, the student should request help. After reading the hint carefully the learner should then decide whether the hint

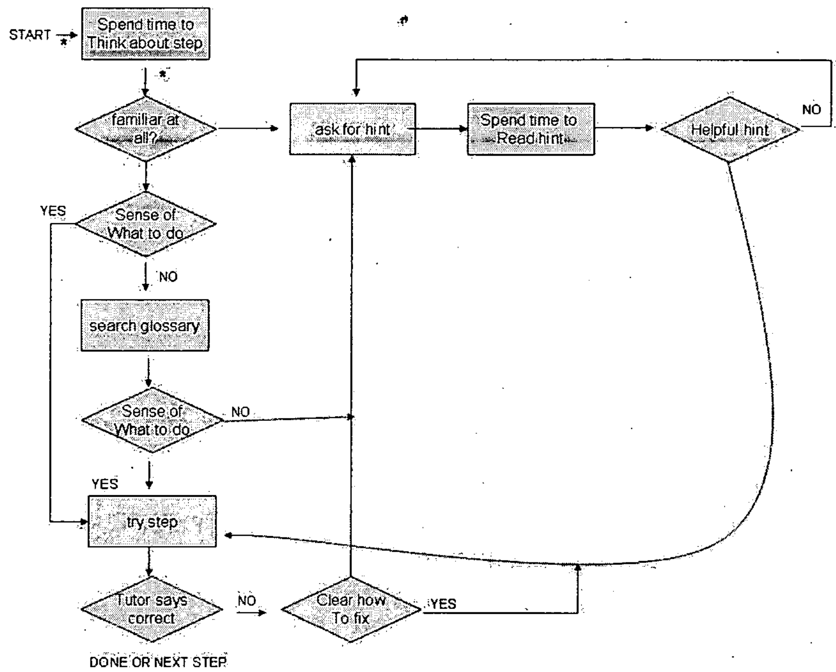


Figure 2.2: A model of good help seeking behaviour (adapted from Aleven et al., 2004)

provides enough information to attempt the step or whether another hint is needed. Compared to classroom models a couple of (social) factors are not taken into account or play a different role. For example, finding a suitable helper is not applicable here as the facilities are predetermined but choosing between the de-contextual help and the more executive-like hint help is a similar action. In addition, some other factors (e.g., fear of embarrassment) have been ignored deliberately.

However, it should be noted that the simplicity of the model lends itself to possible limitations. First of all, some of the steps in the model are computationally difficult to implement. For example, whether the student is familiar with the step or has a sense of what to do is quite abstract and requires assumptions behind their meaning that are not easy to implement objectively. Moreover, as Aleven et al. (2004) also recognise, implementing the model and allowing it to influence interventions would result in feedback being provided very often interrupting students' work. In addition, some of the discussion in the previous section seems not to play a role in the model. It was explained that many different aspects should be taken into account for even the simple act of denying help to the student. The exact point of when to intervene is a difficult issue that has been debated in the field of ITS from its early beginnings (Lewis and Anderson, 1985; Merrill et al., 1995). Given the discussion in the previous section about emotions and motivation, several decisions in the flowchart in Figure 2.4.4 could be improved if the decision steps could be adapted based on students' affective and motivational state as well as other situational aspects.

Therefore, before actually implementing this model (or any similar model for that matter), and in order to improve its effectiveness, apart from more empirical research in student's help-seeking behaviour, the diagnosis of factors that influence help-seeking becomes paramount. This thesis contributes towards the goals.

2.5 Dealing with emotions and motivation in ILE

The aforementioned establish not only the fact that emotions and motivation are important, but also that their detection and consideration during teaching is not necessarily a straightforward task. The discussion in the previous section was narrowed to help-seeking behaviour as one of the most important aspects of students' behaviour in ILE. However, the rough picture painted, both around the issue of emotions and motivation in learning, as well as their links to students' help-seeking, underscores the complex nature of the phenomenon.

Realising the above, researchers in the field of AIED are attempting to improve ILEs by adding components that take into account students' emotions and motivation. The emerging area of 'Affective Computing' (Picard, 1997) has played an important role in this. "Affective Computing is computing that relates to, arises from, or deliberately influences emotion or other affective phenomena" (Picard, 1997) where 'affective' is used in a broad sense to refer to anything pertaining to emotions and motivation. Another term often used is 'emotion-oriented computing' (Schroder and Cowie, 2006).

The applications of this multidisciplinary area are broad and not only related to educational computing. In the context of education, researchers' goals were summarised in the Introduction by what Conati et al. (2005) coined as an attempt to engage educational systems and learners in an 'affective loop'. This involves

- the detection of a learner's emotions
- selection of tutor actions that are beneficial to their learning and
- the synthesis of emotional expressions which would engage the learner in a more 'naturalistic interaction'

Although all three areas are important, the current research revolves around the two first aspects of the above process, and particularly that of *detection*. The issue of selecting appropriate tutor actions has been investigated in detail in the context of ILEs from other researchers (e.g., Malone, 1981; Lepper et al., 1993; del Soldato and du Boulay, 1995; Rebolledo et al., 2006). Of particular interest is the work of del Soldato (1993) who highlighted that planning should be a combination of motivational and instructional planning. According to the authors, the two could often be in conflict and the architecture of the ITS needs to take this into account (del Soldato and du Boulay, 1995). In another work, Rebolledo-Mendez (2007) expanding the work of del Soldato, employs an animated agent capable of addressing various degrees of motivation on behalf of the student.

Regardless of the exact approach followed, the work cited above requires detection of students' affective states, which is one of the most difficult problems in the area. It requires knowledge which even human tutors do not always have. The Introduction mentioned that some researchers have investigated the issue of diagnosis based on facial expressions, voice, or other bodily measures (e.g., Picard and Scheirer, 2001; Kapoor and Picard, 2005; Messom et al., 2005; Litman, 2006; D'Mello and Graesser, 2007). Such methods tend to be very expensive, obtrusive and could interfere with the actual learning process.

Another approach, the one favoured in this thesis, is the use of observable student actions as a means of modelling. Again, as mentioned in the Introduction, some researchers have investigated the issue based on intuition or theoretical models. Methodologically, this approach is often desirable especially if the focus of the research is to investigate other matters. For example, in del Soldato and du Boulay (1995) and Rebolledo et al. (2006) the focus is on motivational planning techniques, and therefore the detection part of the process is only an intermediate to be able to build as research prototypes.

The theory-driven approach is also useful but its accuracy and the extent to which a learning or motivational theory is applicable in the context of ILEs is questionable (du Boulay and Luckin, 2001). For example, many systems employ the OCC model of emotions (Ortony et al., 1988). OCC is the most popular psychological model of emotions in the field of AIED and ITS as it is computationally feasible. It provides a classification scheme for common emotion labels and explains their origins by describing the cognitive processes that elicit them. In OCC emotions are positive or negative reactions to situations and depend on the situation being desirable or not. There are several ways to implement the OCC model; for example, Jaques and Viccari (2007) employ the reasoning capacity of the 'belief-desire-intention' (BDI) approach (Bratman, 1990) which describes an agent as an intentional system.

In addition, researchers in the field recognise the highly complex nature of student modelling and often employ probabilistic modelling and reasoning frameworks to handle the inherent uncertainty of the task at hand (Conati et al., 1997; Mayo and Mitrovic, 2001; Zukerman and Albrecht, 2001; Jameson, 1996). For example, based again on the OCC model, Conati (2002) employs Dynamic Decision Networks (Russell and Norvig, 1995) in order to take into account various pieces of evidence for the emotion detection. Similarly, Morales et al. (2006) also employ the OCC model, but the reasoning behind the system is based on the Dempster-Shafer Theory (Shafer, 1976), which enables the accumulation of evidence along different dimensions for a particular belief of the system.

However, regardless of the exact theory behind the computational approach or the method to deal with uncertainty, the problem of drawing the actual evidence remains. In most of the approaches the types of the exact evidence that is taken into account are based on intuition. One of the most influential paradigms that tries to alleviate this problem is the work of de Vicente (2003) which employs a traditional knowledge elicitation from experts approach. De Vicente elicited rules for recognising a learner's motivational state by asking human tutors to infer students motivational state by watching a recorded interaction of the student with a mock ITS. The analysis of the collected data was mainly qualitative (i.e., the rules were derived based on the tutor interviews). Apart from the exact rules derived (that were perhaps a bit specific to the context) de Vicente's work provides an interesting methodology and establishes an innovative approach to eliciting motivation diagnosis rules beyond intuition or relevant theories.

Another related research project, of particular relevance to the work described in the current thesis, is the work of Morales et al. (2006). The learner model component described there is part of an ILE for teaching mathematics called ActiveMath which is being developed in the context of an EU-funded project called LeActiveMath (LeAM, 2003) (or LeAM for short). The Extended Learner Modeller (xLM), as it is called, comprises of two parts, a Situational Model, responsible for diagnosing short term motivational states using evidence from the students' interaction with the ILE, and the Learner Model which is responsible for the accumulative motivational and affective dispositions of the learner towards the subject domain. More specifically, the Situational Model is an extension of work from Porayska-Pomsta (2003) where an attempt is made to operationalise the theory of Brown and Levinson (1987) mentioned in Section 2.4.1. In particular, autonomy is seen as the need of students to be allowed the freedom to discover knowledge by themselves, and approval as students' needs to have their motivation and emotional addressed by appropriate feedback from the tutor. The situational context is then defined as a combination of factors that impact on the two dimensions of face. This leads to a prototype implementation which employs Bayesian networks to combine the influence of the situational factors.

However, once again, the issue of the actual diagnosis and inference of specific values for the relevant variables of the model becomes important. In the context of the EU project, practical questions needed to be addressed before implementing the model. As one of the deliverables of the project describes, "the exact set of situational factors and their possible values relevant to the domain of mathematics, as well as the manner in which they combine with one another needed to be established in order to enable the situation model to make appropriate calculations" (Andres et al., 2005). To establish the above, the project employed descriptive and qualitative analysis of data collected from tutors when they engaged in a computer-mediated tutorial dialogue with students. The data from this same study are also used in Chapter 6 of the current thesis to investigate the potential of machine learning techniques to derive predictive models of students' affect. Although the exact details of the study are presented in Chapter 6 the relevant findings from the descriptive and qualitative analysis are summarised below as they shaped the understanding of various parts of this thesis. For more details the reader is referred to (Andres et al., 2005; Porayska-Pomsta et al., 2008).

First of all, a general observation was that overall the tutors were able to identify the motivational states of the students despite the fact that their interaction was through a chat interface which limited the bandwidth of information available to the tutor. In addition, some of these factors played a role in the tutor's actions. From a detailed descriptive analysis the most relevant factors identified were (1) student confidence, (2) student interest, (3) student effort, (4) correctness of student answer, (5) difficulty of material, (6) importance of material, and (7) student aptitude. By using Principal Components Analysis the factors were grouped in two sets. The first group includes aptitude, confidence, interest and effort, while the second correctness of student answer, difficulty of material and importance of material. The groupings of factors obtained lend themselves naturally to their different sources of diagnosis. The first group represents the factors whose values need to be diagnosed based on the information obtained on an ongoing basis from the interaction between the tutor and the student and from the student's observable behaviour. The values of the second group are typically obtained from the ILE (e.g., content metadata). Further to that, Andres et al. (2005) describes how the verbal protocols were analysed to obtain insight into the cues from student behaviour on which tutors rely during the interaction. Seven main sources were found: hesitation, linguistic cues, student's achievement level, difficulty of material, spontaneous admissions, granularity of solution steps and student initiative. These sources contribute to the diagnosis of each individual variable based on a set of rules that are derived through interviews, post-task walkthroughs but also unavoidably, possibly the intuition of the researchers.

Once again, when it comes to the exact details of the implementation unavoidably some subjectivity will exist. This is where the work in this thesis finds its relevance by investigating how machine learning techniques can increase the validity of qualitative results.

Chapter 3

Tools and Techniques

3.1 Introduction

This chapter describes the tools and techniques employed throughout the thesis. As the research presented here revolves around issues related to ILEs and their actual use, access to an ILE that is integrated in realistic situations was needed. As the author of this thesis was the main developer of a web-based environment, built specifically for the School of Mathematics of the University of Edinburgh, access to the code and the context where WALLIS was employed was easily established. The fact that the system was to be introduced in an actual educational situation while the current research would be on going, provided the advantage of a real context where the effectiveness of the system overall could be tested. Section 3.2 presents the system, the methodology and theories behind its development. The section is adapted from the following papers (Mavrikis and Maciocia, 2003a, 2002, 2003b; Mavrikis, 2004). Section 3.3 is adapted from Mavrikis (2005) and describes the logging functionalities added to early versions of the system to address the needs of the current research for recording as realistic student-system interactions as possible.

Section 3.4 outlines the machine learning and data mining techniques used throughout the thesis in order to avoid repetition in the following chapters.

3.2 WaLLiS: a Web-based ILE for Teaching and Learning Mathematics

WALLIS (named after James Wallis; the 17th century mathematician) is an ILE developed initially during a KETF[1]-funded project of SHEFC[2] and later supported by funding from the University of Edinburgh. Other parts of the system (in particular the feedback and logging mechanisms) were specifically developed to facilitate the research described in this thesis. Mavrikis and Maciocia (2002) describe the main rational behind the system. In brief, the web-based environment attempts to address the growing concern of researchers and university teachers about the evident problem of the deficiency in mathematical skills amongst science and engineering students (LTSN, 2000; Hunt and Lawson, 1996). Students' diverse backgrounds and the fact that many of them fail to recognise the importance of mathematics for their main degree, make additional support (tutorials, formative assessment etc.) difficult and, in conjunction with the increased intake of students, time consuming. By providing material and employing WALLIS, the School of Mathematics provides additional support to students.

As discussed in the previous chapter, all aspects of the context play an important role in the thesis. Therefore, it is important to describe first the methodologies and theories which influenced the development of WALLIS as well as the system itself.

3.2.1 Theories influencing the development of WALLIS

Following a constructivist point of view and recognising that learners must remain "involved, active and challenged to think and learn about the presented material" (Woolf et al., 2001), WALLIS attempts to bypass the drawbacks of passive learning and the teacher-oriented instructivist point of view that dominates the teaching and learning of mathematics in higher education. As a recent review describes, "the unspoken assumption [...] is that delivery of the content results in learning of the material, through a process of osmosis" (Philips, 2005). This assumption ignores the cognitive and affective processes which lead to learning. Students have to be provided with stimulating

[1] Knowledge Economy Task Force
[2] Scottish Higher Education Funding Council

learning opportunities that will foster self-guided learning, make them play an active role, and be responsible for their own learning process.

Therefore, apart from static theory and example content, that is often provided in this kind of systems, WALLIS provides interactive and exploratory activities. These allow students to freely explore aspects that cannot be covered in static pages, as well as other learning opportunities through interactive exercises. The development of these interactive exercises is strongly influenced by AIED and therefore fits the context of this research particularly well. One of the theories which addresses such issues to some extent and influences the development of WALLIS is contingent instruction (Wood, 2001) and consequently scaffolding; notions proposed to describe exactly the need for a balance between the children's capacity to selectively ask for help and the teacher's effort to take actions, contingent upon activities of the individual learner. This is achieved by recruiting student's interest, establishing and maintaining an orientation towards task relevant goals, demonstrating how to achieve goals and helping control frustration ensuring that the student is neither left to struggle alone nor given too little scope for involvement and initiative in the task (see Wood, 2001). An important issue that arises when one tries to teach in such a way, is to know when and how much help a student needs to complete a task. As described in the Background, a common solution in ITS s to employ the learner's use of help-seeking to influence the tutoring process. However, since many factors influence students' help-seeking behaviour more research is needed to address this issue.

The development of WALLIS was inspired by methodologies that call for "continuous refinement of system behaviour" (Woolf et al., 2001) through careful user studies (Koedinger, 2001; du Boulay and Luckin, 2001). As described in the Background, a methodology that takes all these considerations into account is Persistent Collaboration Methodology (PCM - Conlon and Pain,1996). Accordingly, the design and development of WALLIS has been stimulated by research in the field. Simultaneously research and development on the system contributes to further research in the field, as PCM expects. On one hand, the research described in this thesis and other related projects (e.g., Mavrikis, 2001; Abela, 2002; Hunn, 2003) have strongly influenced its design. On the other hand, the initial decision to investigate the issues in this thesis was based on early observation phases of the prototype versions of the system and how it was integrated in the educational situation. Since its first days, WALLIS has been used both as a medium that supports teaching in the School of Mathematics as well as a research tool that influences theories on interactive learning and the system's further development. In the spirit of PCM, further cycles of observation, reflecting and redesign will help improve WALLIS while contributing to the 'wheel' of knowledge about ILEs and AI techniques and tools.

Through the iterative phases and over the years of its use the feedback mechanism was adapted, more misconceptions were targeted, and aspects of the system related to HCI were improved based on interviews and observation of students working under controlled conditions. All these resulted in a stable system that students use under realistic conditions and for which there is evidence that they are learning from it. This enabled the investigation of certain issues pertaining to the current research materials that would otherwise be difficult.

The following section describes the architecture of WALLIS, the material that are available and the feedback it provides. This description is constrained to the features related to this thesis. For more technical details the reader is referred to Mavrikis and Maciocia (2002, 2003b, 2006).

3.2.2 The architecture and environment of WALLIS

WALLIS, is a web-based environment that hosts contents which includes pages of theory or examples that present the material as well as interactive web pages that involve interactive exercises. The exercises comprise activities either in the form of applets, during which students interact with a microworld (in the sense of Balacheff and Sutherland, 1993) called DANTE (Mavrikis, 2001, 2004) exploring a concept, or other exercises which are more constrained, such as cloze or multiple choice questions with multiple steps, which are delivered as a web page with interactive elements.

The overall environment of WALLIS is depicted in Figure 3.1. It follows a design that is similar to many state-of-the-art eLearning environments (Moodle[3],WebCT[4]) but

[3]http://www.moodle.org

[4]http://www.webct.com

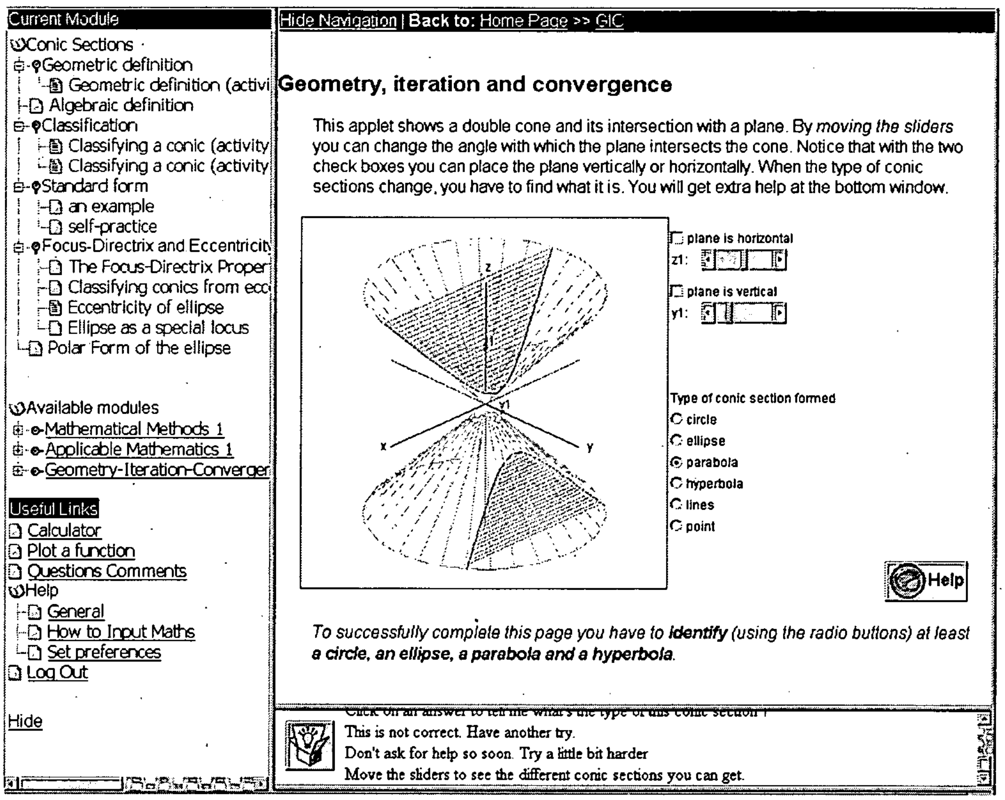


Figure 3.1: The overall environment of WALLIS

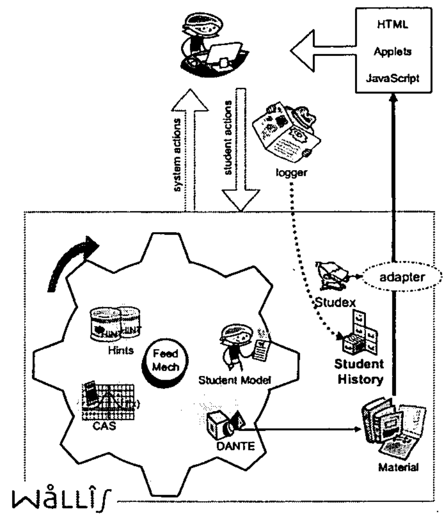


Figure 3.2: The architecture of WALLIS and how its components communicate (adapted from Mavrikis and Maciocia (2003b))

belongs to what the Advanced Distributed Learning initiative[5] calls advanced eLearning environments (ADL, 2001) in the sense that it combines features of content-based approaches with adaptive educational strategies from ITS (see Section 2.1 where this difference is outlined). Accordingly, apart from the usual components of the system that deliver the material (the main central frame in Figure 3.1) and the tree-like map of the content (typical in many eLearning systems), WALLIS incorporates a feedback frame at the bottom of the window where all feedback is delivered to the students (see Figure 3.1).

In simple terms, architecturally WALLIS is separated into client and server components that are responsible for the delivery of content and its adaptation as well as the feedback that the system provides. Its architecture is presented in Figure 3.2 and is described in the following subsections.

3.2.2.1 Contents and feedback in WALLIS

WALLIS addresses various concepts such as functions (function domain, odd and even

[5] http://www.adlnet.org

functions, slope and gradient for linear functions etc.), differentiation, integration and vectors by presenting the corresponding theory pages and employing (where appropriate and feasible) exploratory activities or interactive feedback-enabled exercises. Particularly to accommodate this research a section of the 'Geometry and Iteration' course that deals with Conic Sections in detail was authored and employed over the years in several different setups (see Appendix B for the material).

3.2.2.2 Navigation

In an attempt to address the common problem in web-based environments of students getting easily confused and lost by many hyperlinks (for examples, see Thuring et al., 1995; Conklin, 1987; Brusilovsky, 1996) links between the actual content of WALLIS are avoided as much as possible. Instead, the navigation frame with the map of the contents changes according to the course students attend and gives them control over which part of the material to study. When a page is visited the tree is annotated accordingly to inform the student (see icons on the left of the links of Figure 3.1). This information is kept in a *Student Model* that the system maintains for each student.

3.2.2.3 Adaptation of material

Because the system is used in a more or less a specific context and in a particular didactical approach, the content is adapted just to student preferences (colours, sizes etc.) and very broad aspects of their profile (e.g., their degree) rather than to different didactical styles or learning scenarios as is the case for other adaptive system, such as ActiveMath (Melis et al., 2001). All these are kept in a component called *Studex* that indexes students' information.

3.2.2.4 Feedback provision

The important adaptive feature of the system is the feedback mechanism which provides feedback and suggestions to the students. As explained in Mavrikis (2001); Mavrikis and Maciocia (2002) initial pilots of the system made clear that pop-up windows were considered quite annoying and, in the particular case of WALLIS, where student actions during the interactive parts would produce feedback interrupting the student's action, this approach was not feasible. In other interactive exercises (see Figure 3.4) delivering feedback in the same page, as usually happens in similar systems (c.f. Melis et al., 2001), was confusing students. Despite the use of colours or other visual indications, the fact that the content of the page was suddenly changing and elements were moving around caused some students to get lost. Therefore, all feedback is delivered in the feedback frame at the bottom of the window (see Figure 3.1). The main purpose of this is to provide feedback without interrupting students' work.

The feedback mechanism is mainly inspired by theories of cognitive skill acquisition such as ACT-R (Anderson, 1993) and cognitive scaffolding (Wood, 2001) and follow similar approaches to the CMU Cognitive Tutors (Anderson et al., 1995). During the activities, help is offered in an instrumental way trying to predict students' misconceptions and provide them with as much help as necessary to progress with the activity, therefore turning a problem in which they do not necessarily have sufficient knowledge to solve, into a teaching aid from which step-by-step they gradually learn by practise.

The mechanism relies on different components for its 'intelligence', depending on the activities. During the exploratory activities, the system relies on an adaptation of the feedback mechanism developed for DANTE (Mavrikis, 2001, 2004) and JavaMath[6]. A threaded mechanism tracks the goals that the author of the activity sets and students have to achieve. The goals of the activity form a tree structure (see Figure 3.3) where goals are comprised of a number of subgoals each of which has a number of completion conditions and misconceptions associated with it.

The goals, depending on the activity, involve selecting an answer from a multiple choice question, putting objects into certain positions, giving numerical answers and so on. The feedback mechanism comprises production rules authored in Java (in old versions) and recently in JESS [7] (see Hunn, 2003; Mavrikis, 2004).

Along with these, there is appropriate feedback associated with each goal, and other messages to be delivered when no action is taken or when they achieve a goal etc.

[6] JavaMath (maths.hws.edu/javamath) is a collection of graphical and other mathematical objects that help present but also validate input, calculate integrals, derivatives etc.

[7] http://herzberg.ca.sandia.gov/jess/

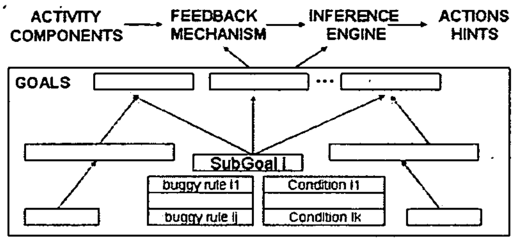


Figure 3.3: The representation of DANTE's goals

More details about the feedback mechanism can be found at Mavrikis (2001, 2004); Hunn and Mavrikis (2004) but in general this follows an incremental hinting process that changes according to the time passed, the amount of help a student requested (incrementally demonstrating the answer if necessary), the current goal but also the student's ability.

For example, at the simple exploratory activity seen at Figure 3.1 students can explore conic sections. By moving the sliders they change the plane's coefficient that intersects with the double cone. Afterwards, they have to choose from a multiple choice question the conic section produced. During their interaction, they receive feedback that helps them explore the activity more efficiently such as hints to move the sliders to different positions, prompts to provide an answer, to rotate the double cone etc. Similarly, when they select an answer (note that this is one of the goals for this activity) students receive feedback based on the preset rules (see Figure 3.1 for a short example). Finally, when a student explicitly asks for help, depending on her current goal and level, the system provides feedback on how to achieve that. For example, *'Think first how many branches the conic section has'* or *'Drag and turn the cone to see it from a different position'* that would potentially, if the student keeps on failing, lead to the system providing an answer.

A similar approach is followed for the interactive activities which are common web-pages that include a form for the interaction (see Figure 3.4), buttons for the student to submit their answers (unless it is a multiple choice question) and request hints or the solution. The feedback mechanism in this case relies on a Computer Algebra

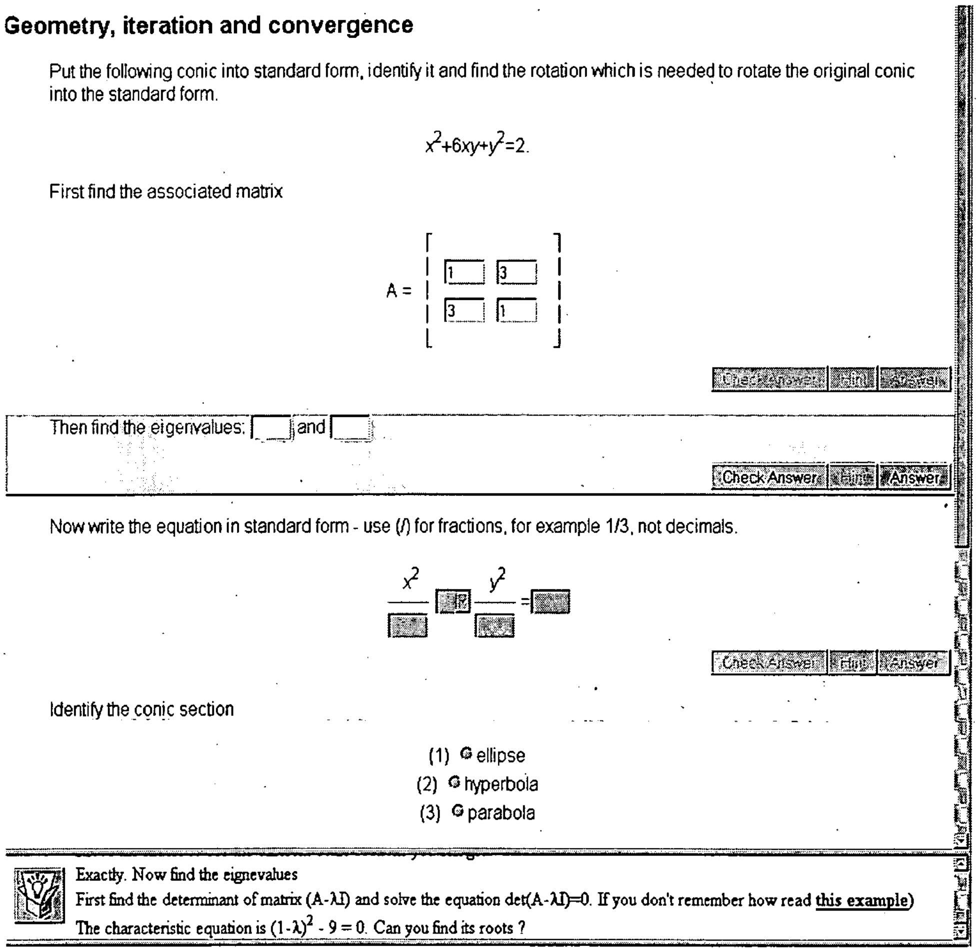


Figure 3.4: A self-practise interactive exercise

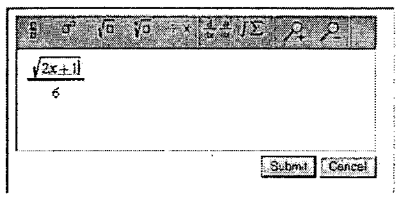
System (CAS)[8]. After submitting an answer, if it looks like one of the common misconceptions encoded in the system previously, then students are provided with corrective feedback (referred to as `feedback_corrective` in the data analysis). If the error is unknown to the system then negative feedback is provided (`feedback_negative`) with a suggestion for the student to try again or to ask for a hint. If the answer is correct, positive feedback (`feedback_positive`) is provided and the student can move on to the next step or (if they wish) request the solution of the previous step (perhaps to elaborate on). An exercise step is usually associated with 3 to 5 help messages (`hint_1 ...5`) and appropriate corrective hints based on predicted misconceptions (`hint_corrective`). The first few hints rephrase the question statement and give instructions to the student on how to proceed in order to come closer to solving the problem. The last hint is very specific, almost providing the answer. After all possible help is provided, if the student is still struggling, they can request the solution which is then provided together with an explanation. If a common misconception is observed during their attempt to solve the problem, the solution can be adapted accordingly to highlight the students' mistake.

Finally, there is another level of feedback which deals with the pages students access and provides suggestions in relation to which page the student should select, based on the exercises and examples they have already covered and the theory pages they have visited and read. If a student is lost on a page they can request a suggestion and if they haven't completed a prerequisite of the page they are interacting with, WALLIS suggests that they visit the relevant page. The mechanism is based on a similar tree structure of goals and subgoals with the one in DANTE, which was described above.

3.2.2.5 Mathematical input, notation and verification

As described in Mavrikis and Maciocia (2002), from early prototypes of WALLIS, it was evident that students faced serious problems typing their answers in a linear format. It was very annoying for most of them, regardless of their level of competence and computer literacy. Similarly, they complained that too much effort was needed to quickly understand the linearly typed mathematics in the feedback frame. This

[8]http://www.maple.com



```
<math> <mrow>
    <mfrac> <mrow> <mrow>
        <msqrt>
            <mrow> <mn>2</mn>
             <mo>&InvisibleTimes</mo>
             <mi>x</mi>
             <mo> + </mo>
              <mn> 1 </mn> </mrow>
        </msqrt> </mrow> </mrow>
        <mrow> <mn>6</mn> </mrow>
    </mfrac> </mrow>
</math>
```

Figure 3.5: Using WebEQ to input sqrt(2*x+1)/6 and its underlying MathML.

was also affecting students' ability to pay attention to the feedback, despite the fact that it was highlighted. Although one could argue that mathematics students should learn to type and understand a linear format, it is surely not the right context to do so. The cognitive load to understand and type in this format (which many mathematical learning environments employ) obstructs their learning of more important aspects at this stage.

In order to address the aforementioned concerns, which are also related to the affect and motivation of the students, the version of WALLIS which is used for the studies in this thesis employs an input editor (see Figure 3.5) and the feedback is provided in a more friendly notation for students. The input editor (WebEQ[9]) provides the ability to transform a student's input to MathML[10] which is then sent to the CAS where the answer is evaluated.

3.2.2.6 Logging students' actions

WALLIS logs some basic student actions (page completion and help requests) in order to populate the *student model* with appropriate information in relation to the students' completion of goals. The component responsible for this is essentially designed as an agent that monitors and records the learner's interactions by sending them to the server for storage (see *logger* in Figure 3.2).

[9]WebEQ provides an equation editor applet (see http://www.dessci.com). For collecting research data a modified version of WebEQ based on WebEQ SDK API adapted mainly to capture events and mouse movements in the editor.

[10]http://www.w3.org/Math

3.3 Logging and Replaying Students' Interaction

For the purposes of this research, the basic recording functionality of WaLLiS was extended. Analysing the log files of the web server would provide limited information and, since a more exploratory approach was followed, it was not possible to make prior decisions about which aspects of students' behaviour are relevant to log. Most of the research in the field tends to do that and log at a coarser granularity. When this fits the intended analysis it leads to very interesting and successful results (for example, Baker et al., 2004a; D'Mello et al., 2006a; Stevens et al., 2004; Arroyo and Woolf, 2005). However, for the research conducted in this thesis there was a need to go a step further and record as much information as possible.

Therefore, the logging mechanism of WALLIS was enhanced with the ability to log remotely *every* aspect of students' actions (from simple button clicks to all mouse movements). This served two purposes. First, it enabled replays of realistic student interactions as the logging mechanism records all actions in a timestamped manner that can be replayed. Second, it enabled detailed replays and analysis of students actions at any desired level. Although using video recording would be sufficient for replaying students' interactions, it would hinder any other analysis as it would require time-consuming coding of the actions that take place. Recording at a low level required only the development of this mechanism. More details about the mechanism are provided in Mavrikis (2005) and in some detail in Appendix E. The replays are conducted employing an adapted version of the system in a local computer and a parser that loops through the recorded, timestamped interaction and replays the interaction of the student as if they were using the system. More details are also provided in Appendix E.

3.4 Educational Data Mining

"Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students, and the settings which they learn in."[11] This section briefly presents the data mining and machine learning techniques

[11] http://www.educationaldatamining.org/index.html

used throughout the thesis for the benefit of the reader who is not familiar with them. Although common statistical (e.g., ANOVA, multiple regression etc.) techniques are also employed throughout the thesis, these are not described in detail as they are established already in the field of education and well known. The following subsections describe *decision trees* (3.4.1), *clustering* (3.4.2), *Bayesian networks* (3.4.3), *feature selection* (3.4.4), and how the learning *accuracies* of the employed algorithms are evaluated (3.4.5). Finally some other mathematical techniques used in the thesis such as the *Mahalanobis distance* and a *discretisation* method are outlined in 3.4.6. For the descriptions that follow Witten and Frank (2005); Bouckaert (2004); Bishop (2006) and other textbooks (e.g., Russell and Norvig, 1995) were influential but are not explicitly cited below for clarity purposes.

For all the machine learning analysis throughout the thesis, the Waikato Environment for Knowledge Analysis (WEKA, Witten and Frank, 2005) is used. WEKA is a data mining platform written in Java. It has a large community of users and includes many of the most traditional machine learning algorithms. In addition it is open source and therefore it is possible to adapt an algorithm or develop a new one for the particular needs of an analysis. For example, for the current thesis, the *discretisation* and *feature selection* filters that are applied to the data prior to some of the machine learning analysis were changed according to the preferred ones as described below.

3.4.1 Decision tree induction

Decision tree induction algorithms are one of the most popular methods of predictive modelling. They are used to produce graphical diagrams in the form of a tree with nodes and branches that essentially represent rules in a hierarchical structure. To draw an example from the current research, in Chapter 6 a decision tree that provides predictions about whether students' confidence is decreasing or increasing was derived from appropriate data. An illustrative, simplified decision tree is presented in Figure 3.6.

The graph comprises nodes (or vertices) connected by links (or edges). Each node represents an attribute and the labels on the edges between nodes indicate the possible values of the parent attribute. Following a path from the root to a leaf creates a rule that shows if confidence is decreasing or increasing given the values of the attributes

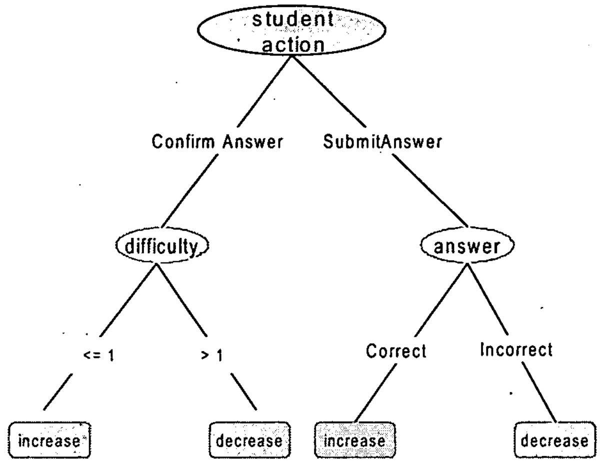


Figure 3.6: Example of a simplified decision tree from Chapter 6

in the path. For example, the fictionary tree in Figure 3.6 indicates that when students are confirming their answer and the value of the variable *difficulty* is larger than one, their confidence decreases.

One of the reasons behind the popularity of decision trees is that they provide rules that are human inspectable and are easier to interpret and to validate. Decision trees can be transformed into rules and hence they can be implemented straightforwardly in a rule-based environment.

Given some data, a decision tree induction algorithm can be used to classify them into nodes or leaves that are as homogeneous as possible with respect to one of the variables (usually referred to as *class*). This classification is defined in terms of attributes, essentially providing a predictive relationship between the attributes and the class. This is the approach followed in this thesis. For example, decision trees similar to Figure 3.6 were constructed (or 'learned') by presenting to the algorithm appropriate instances which contained student actions and tutor's diagnosis of students' confidence.

WEKA provides several decision tree induction algorithms. Throughout this thesis J4.8 is used. J4.8 is based on a slightly improved version of the popular decision tree algorithm C4.5 of Quinlan (1993), and specifically Revision 8. The algorithm operates over a set of instances and generates a decision tree by selecting the attribute that separates the classes best and places it at the root node. The selection is based on the information value or entropy of each attribute (for more details see Witten and Frank, 2005; Quinlan, 1993, 1996). A branch is created for each possible value of the attribute, partitioning the example space into subsets to which the algorithm is applied recursively until all instances of a node have consistent classification.

3.4.2 Clustering

Clustering is a data mining technique used to create categories that fit observations. A clustering algorithm searches groups of examples that belong together. In contrast to the decision trees induction, clustering techniques are useful when there is no class to be predicted but rather when the instances are to be divided into natural groups. The assumption when interpreting the results is that the clusters formed reflect some properties of the instances, which cause some of them to be more closely related to each other, than they are to the instances in another cluster. To draw an example from this thesis, in Chapter 4, presents an attempt to determine groups of students that behave similarly and derive student types in terms of their behaviours with the system. A number of variables, such as help frequency, time spent per hint etc. were used to characterise students' sessions with the assumptions that it would be possible to design appropriate interventions for students who manifest similar behaviour.

There are two general types of clustering methods: nonhierarchical and hierarchical. Nonhierarchical methods require that the number of clusters is known in advance. However, without detailed knowledge and insight to the structure of the dataset, it is difficult to determine the number of clusters in advance objectively. Hierarchical clustering algorithms approach the problem without a predetermined number of clusters. The clusters are derived after a series of incremental steps which begin either by considering all instances in one cluster or treating each instance as its own cluster. In the first case the algorithm divides the clusters of each step (partitioning methods) and in the second it incrementally merges clusters (agglomerative methods).

The particular algorithm employed in this thesis (COBWEB as implemented in WEKA) follows a hierarchical approach and was preferred as it is the less subjective (even some partinioning methods require a number by which to divide the initial clusters). Its incremental nature allows clustering of new data to be made without having to repeat the clustering already made. It starts with a tree consisting of just the root node, from there instances are added one by one, with the tree updated at each stage. Updating may involve (a) finding the right place to put a leaf representing the new instance, or (b) radically restructuring the part of the tree that is affected by the new instance by either (i) creating a new class containing the instance, or (ii) merging two clusters to include the new tree, or (iii) splitting a cluster in order to accommodate the new instance. A quantity called *category utility* measures the overall quality of a partition of instances into clusters and facilitates the decision above (for more details see Witten and Frank, 2005, pp.260-262).

3.4.3 Bayesian networks

Bayesian networks are a special case of a wider class of statistical models called graphical models which offer a theoretically well-founded way of representing probability distributions in a graphical manner. Bayesian networks, in particular, are directed graphs that represent a set of variables (nodes) and their probabilistic dependencies. They are drawn as a network of nodes, one for each attribute, connected by directed edges in such a way that there are no cycles. For example, the graph shown in Figure 3.7 illustrates a Bayesian network, which, in the research presented in Chapter 5, was used to predict whether a student needs help or not in a certain question.

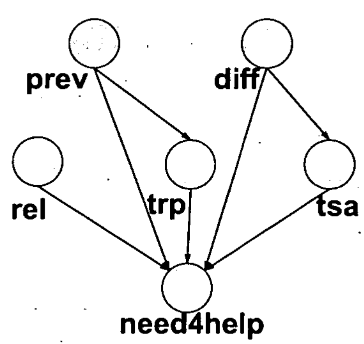


Figure 3.7: Bayesian network example from Chapter 5

In Section 2.5, it was mentioned that the highly complex and uncertain nature of the information about students is recognised more and more in the field of AIEd and ITS. To address this uncertainty, probabilistic frameworks are often employed behind the reasoning process of the system and Bayesian networks are used to model various aspects needed in the system, such as modelling the domain (e.g., Gertner et al., 1998; Murray and VanLehn, 2000), the student's mastery of knowledge items (e.g., Reye, 1998), or other relationships between observed student actions, student internal states and outcomes (Mayo and Mitrovic, 2001).

Similar to expert systems or rule-based approaches, Bayesian networks are often constructed by experts intuitively or based on some theory or empirical data. Experts can specify the complete structure of the network and/or the conditional probabilities. Apart from the difficulties in doing this, there is often an issue of subjectivity and bias introduced in the model, similar to any other knowledge engineering process. However, it is possible to learn both the structures and the conditional probabilities of the network from data. It is the latter that is of interest in this thesis.

In WEKA, learning a Bayesian network is considered as a learning task of finding an appropriate classifier for a given dataset with a class variable and a vector of attributes (Bouckaert, 2004). The learning is a two stage process of first finding an appropriate network structure and then learning the probability tables. There are several algorithms for learning the structure of the network. On the one hand, learning a network structure can be approached as an attempt to optimise a scoring function that measures the quality of the network structure (this is referred to as *local score based structure learning*). On the other hand, the problem can be considered as a task of learning a network structure that represents the independencies in the distribution that generated the data (this is referred to as *Conditional independence test based structure learning*).

For local score based structure learning, one of the best options available when considering learning speed is K2 (Cooper and Herskovits, 1992). In this algorithm, nodes representing attributes are arranged with edges interconnecting them. The algorithm follows a greedy approach, during which each node is reconnected to previously visited ones in an attempt to maximise the overall score of the network. To avoid localisation and overfitting, a maximum number of parents is set for every node. In addition, to avoid cycles while traversing the graph, the ordering is predetermined and only previously traversed nodes are considered in the optimisation process. However, this can be a constraint as it introduces a dependency between the result and the initial ordering of the attributes. In addition, as is the case with greedy algorithms, the resulting network represents a local maximum of the scoring function. Therefore, running such algorithms several times with different random initial configurations can yield better results. Another approach considered in this thesis is called Hill Climbing (Buntine, 1996) which is similar to K2 but adds and removes arcs with no fixed ordering of the variables. For other approaches, Witten and Frank (2005); Bouckaert (2004) provide detailed information.

In this thesis, conditional independence test based structure learning methods were preferred as they stem from the need to uncover causal structure in the data. Although directed edges in a network do not necessarily represent causal effects, by properly representing the conditional independences in the data, these methods attempt to learn causality. The ICS algorithm (Verma and Pearl, 1992) as implemented in WEKA starts from a complete undirected graph and tries to find conditional independencies in the data. For each pair of nodes, it considers subsets of nodes that are neighbours to the pair. If an independence is identified, the edge between the pair is removed from the network structure and the arrows are directed accordingly (i.e., from each node of the pair to the node that justified the removal of the link). In order to direct any remaining arrows, common sense graphical rules are applied (see, Verma and Pearl, 1992, for details).

3.4.4 Feature selection

Throughout the research conducted here, there was often a need to select appropriate features (attributes) from a set of features that are empirically derived before presenting them to any learning algorithm. For example, in Section 5.3.1 a model was developed that provides a prediction of the ability of students to answer a question correctly without any need for help. This model was learned from data that include features such as the *time spent on related page*, *time spent on attempt*, *difficulty of the item* and *the* *type of the answer required* (see Table 5.1). Because these features are empirically derived some of them (in this case the type of the answer) may be correlated with other affecting the accuracy of the result of the learning algorithm.

As Yu and Huan (2003) describe, high dimensional data with irrelevant and redundant information may degrade the performance of learning algorithms instead of helping them come up with an accurate result. The feature selection literature shows that, apart from the obvious improvements in speed (an issue which is not very relevant in the analyses presented here as all learning is conducted offline), along with irrelevant features, redundant features also affect the accuracy of learning algorithms (Blum and Langley, 1997; Hall, 2000). Therefore, by eliminating them while learning, more accurate models are derived. In addition, feature selection has also the potential to enhance the comprehensibility of the result, since it is usually more simple (for reviews see Blum and Langley, 1997; Kohavi and John, 1997; Yu and Huan, 2003).

The feature selection technique used throughout the thesis is called Fast Correlation Based Filtering (Yu and Huan, 2003) and its investigation was inspired by its successful use in similar research in the field by Baker (2005) who employs it to reduce the features available to the algorithm while automatically searching possible models to fit the needs of his research (for more details see, Baker, 2005).

3.4.5 Evaluating machine learning outcomes

The machine learning tasks addressed in this thesis can be approached primarily as classification tasks and therefore, their performance is measured in terms of their accuracy in predicting the class in a test set. However, when large amounts of data are not available, a technique called *repeated cross-validation* is commonly used to deal with the problem. Briefly, the technique partitions the data into a fixed number of folds and each fold is used for testing while the rest of the data are used for training. This process is repeated in a way which guarantees that each class is properly represented in both training and test sets. The latter is often reffered to as *stratification* and therefore the method as stratified x-fold cross-validation. Finally, it is also customary to repeat this procedure a number of times and average the results. This produces a more reliable error estimate. This is is the preferred technique employed throughout the thesis.

Another way to measure the reliability of the result is the Kappa statistic (Cohen, 1960). The Kappa statistic was first introduced as a means of measuring agreement between two observers of a psychological behaviour while taking into account the agreements that would occur by chance. In data mining, it is used to measure the agreement between the categorisations of the classifier and the observed ones of the dataset. Values of Kappa close to 0 demonstrate that the agreement could be attributed to chance. Values closer to 1 demonstrate that these classifications would not occur by chance.

In addition to these evaluations, another aspect to consider when evaluating is the cost of making wrong decisions. WEKA provides particularly useful measures often used in the information retrieval field, the *precision* (or specificity) and *recall* (or sensitivity) of an algorithm which are combined into what is referred to as *F-measure*. Recall is of particular importance in algorthims that classify educational data as it measures the number of instances classified correctly over the total number of instances in this class. The larger its value is, the less false negatives (positive instances that are incorrectly classified) the algorithm produces. These were taken into account when deciding on the accuracy of the machine learning outputs in this thesis.

3.4.6 Other Mathematical Techniques

3.4.6.1 Mahalanobis distance

Mahalanobis distance (Mahalanobis, 1936) is often used in discriminant analysis to detect outliers. It is used as a metric to test whether a particular instance would be considered an outlier relative to a set of group data. In Chapter 4 it is used for the distance measure of the clustering algorithm and throughout the thesis it is used in the place of the traditional Euclidean distance, when there is a need to measure the distance between elements of whole vectors. The two distances are similar to the difference that the Mahalanobis utilises group means and variances for each variable as well as the correlations and covariance of the data set. It is also scale-invariant (Maesschalck et al., 2000).

Formally, the distance of a vector $x = (x_1, x_2, \ldots, x_n)^T$ from another vector, $y =$

$(y_1, y_2, \ldots, y_n)^T$ is defined as

$$D(x) = \sqrt{(x-y)^T \Sigma^{-1} (x-y)}$$

where Σ is the covariance matrix.

3.4.6.2 Discretization

It is often the case that the machine learning algorithms described above can only handle or perform well with categorical attributes while some of the attributes that appear in the data throughout the thesis are continuous. In these cases they should be discretized into distinct ranges. Even when algorithms can deal with numeric attributes it is often better to discretize them to ranges that have some semantic meaning and employ information about the attribute itself. Otherwise the algorithm (usually the decision trees) tends to discretize the variable in different cut-off points making the result less human-inspectable and less useful in terms of implementation for the system.

Therefore, continuous variables such as *time* or *frequencies* (e.g., frequency of help-seeking) , are discretized prior to any machine learning analysis. The number of breakpoints was chosen empirically in an attempt to maintain the disproporitonallity of a normal distribution and the notion of the fuzzy linguistic variables. They were also inspired by similar discretisation techniques in time series data mining (e.g., Lin et al., 2003) where the need to discretize time series is lined to dividing a Gaussian distribution into an arbitrary number of equiprobable regions. That is breakpoints are a sorted list of α numbers $B = \beta_0, \beta_1, \ldots, \beta_{\alpha-1}, \beta_\alpha$ (with $\beta_0 = -\infty$ and $\beta_\alpha = \infty$) such that the area under a N(0,1) Gaussian curve from i to $i+1$ is $1/\alpha$. These breakpoints can be determined by a statistical table. Accordingly, if three regions are needed the breakpoints are $z <= -0.84$ for `Low`, $-0.84 < z < 0.84$ for `Medium` and $z > 0.84$ for `High`. For five regions the breakpoints are $z <= -1.28$ for `Very Low`, $-1.28 < z < -0.52$ for `Low`, $-0.52 <= z <= 0.52$ for `Medium`, $0.52 < z < 0.84$ for `High` and $z >= 1.28$ for `Very High`.

3.5 AI tools

Although Weka's code is reusable and could be used for tasks beyond just learning classifiers, it is not optimised for inference when a model is already developed. Two tools from the AI community are used here for the purposes of inference. The first is a rule-based one called *Java Expert System Shell* or *Jess* [12] and is already employed in the feedback mechanism of WALLIS as described in Section 3.2.2.4. Its relevance in this thesis is that outcomes from decision trees can be transformed straightforwardly to JESS rules enhancing the system's intelligence.

The second tool, called *JavaBayes*[13], is a platform for creating and manipulating Bayesian networks. Its inference engine can be used independently of the rest of JavaBayes in other systems. As its name suggests, it is also written in Java which makes it easy to link with the rest of the components of WALLIS. For the purposed of this thesis, the outputs of the Bayesian models from WEKA that are developed in Chapter 5 are saved in Bayesian Interchange Format (BIF) and are loaded into JavaBayes for inference. Since this is done once, the models are saved manually at this stage but it is possible to automatise the process in the future. The inference engine is described in detail in Cozman (2000).

[12] http://herzberg.ca.sandia.gov/jess/
[13] http://www.cs.cmu.edu/ javabayes/

Chapter 4

Patterns of students' behaviours

4.1 Introduction

As Section 2.4.2 established, the way students behave in any educational situation plays a major role in determining how the situation will unfold and can has an impact on their learning. In particular in ILEs, where the system, compared to a face-to-face interaction, has access to a limited bandwidth of information about students, their actions are even more important. However, as Aleven et al. (2003) also highlight, more research is needed on the ways students interact with educational software and which of them are conducive to learning. Little is known about the exact interaction students have with the system, particularly when they are working in their own time and location. This lack of knowledge often leads to designs which are based on intuitions about the way they are used and the features they should contain. The purpose of this chapter is to provide a better understanding of students' behaviours in WALLIS (and consequently other similar ILEs).

As already mentioned in the Background, several factors influence students behaviours and particular help-seeking. Therefore, the first question that this chapter seeks to address is:

- Which of the factors that influence students' help-seeking behaviours are relevant to the context of WALLIS and what are their implications for the current research?

Having a better understanding of these factors an in-depth exploratory analysis was performed with the following goals:

- identify usage patterns, and
- determine relationships between measures of student-system interaction and learning.

As Section 2.2 discusses, there are several ways to analyse students' interactions. The approach followed here was partially qualitative. It was based on *unobtrusive observations* of students' interactions in WALLIS which provide a clear picture of how students interact with the system. Based on the ecological validity principle driving the studies in this thesis (see Section 2.2), the analysis was performed on students' realistic interactions when they work over long periods with the system and in their own time and location. In order to explore some of the aspects observed in more detail, students' semi-structured interviews complement the otherwise exploratory analysis which is performed employing data mining (specifically clustering) and statistical techniques (differences of group means, multiple regression and correlation analysis). Section 4.2 presents the various studies and the data-sets that facilitated the analyses here.

The rest of the chapter is divided as follows: Section 4.3 discusses the fact that familiarity with the system, the exact content provided and task-orientation are particular important in the context of WALLIS. In particular, by comparing data from the usage of the system in classroom, the influence of *task orientation* and teacher expectations was investigated in more detail. This had implications for the design of subsequent studies.

Section 4.4 outlines an attempt to derive a number of student types in terms of their behaviours with the system. Even if the initial expectation of deriving a small number of groups was not met, patterns of system usage and interactions between variables emerged that were used in the rest of the analysis. Section 4.5 describes in more detail the identified behaviours and their relation to learning, drawing implications for a redesign of the system, and confirming or challenging (based on data) some of the intuitive decisions that are often taken when designing ILEs. Finally, Section 4.6 summarises the most important findings and the implication for the chapters that follow.

4.2 WALLIS courses and data-sets

The following sections present the analysis of a number of data-sets. Their collection was possible thanks to the iterative design methodology behind the WALLIS project (as described in Section 3.2.1) and the integration of the ILE in the teaching and learning of the School of Mathematics of the University of Edinburgh. The courses where WALLIS is mostly used and of which data are analysed later are *Mathematical Methods* and *Applicable Mathematics* (referred to as `MM/AM`), and *Geometry, Iteration and Convergence* (referred to as `GIC`). The purposes of the data collection vary and were often revisited with different purposes in mind to establish some of the related claims. These are outlined below. A more detailed chronology of the studies and data collection that guided this thesis overall is also included in Appendix A.

4.2.1 Mathematical Methods/Applicable Mathematics

Mathematical Methods (`MM`) and Applicable Mathematics (`AM`) are courses used for what is often referred to as *Service Mathematics* and are taught to first year engineering students as additional support for their studies. As mentioned in Section 3.2.1 WALLIS was initially designed for these courses and particularly to allow students to revise and tackle gaps in their knowledge before their exams.

4.2.2 Geometry Iteration and Convergence

Geometry Iteration and Convergence `GIC` is a second year module undertaken by honour students. With the lecturers' agreement, the course was used deliberately as a means of conducting studies for this research. Materials were built for one of the last concepts taught in this module; *conic sections* (see Appendix B). The reasons for choosing this particular course and concept are explained below.

First of all, the materials taught were unknown to the students and they constitute a rather individual unit. While some students had some contact with the subject before, this was very superficial. Apart from the lecturer's experience, this fact was also established from a pre-test during the first pilot as well as a study described in Chapter 6. In addition, with the agreement of the lecturer who was teaching the course, it was possible to deliver the material solely through WALLIS rather than in conjunction with classroom teaching. Moreover, one of the activities (that of converting a conic section into its standard form), which was used for most of the analyses in relation to help-seeking and performance, was presented using a different methodology from most available online materials or textbooks. This helped to establish that any performance results are reasonably (if not solely) attributed to the students' interaction with the system and not other external factors. In particular, with the collaboration of the lecturer, a particular question on the students' final exam was designed to specifically test long-term knowledge retention. Finally, it was possible to establish metrics of previous knowledge of the particular concepts that are prerequisite to understanding the material in the system. The course is taught using formative assessments through tutorials that the students have to attend. Their performance in these tutorials together with their performance at a question in a prerequisite course *Solving Equations* (SEQ) were standardised and averaged to form a measure of previous knowledge[1].

4.2.3 Data-sets

The context, and the goal behind the associated data collections and their analysis presented in this chapter varied over the years (2002-2004). The differences, and the main goals behind each one, are presented below. They are identified throughout this chapter, and in other relevant parts of the thesis, by the course title and the academic year the system was used.

- MM/AM02

  The application of the system for the MM and AM courses is not done in some directed way. There are no activities that the students *have* to complete. However, students are given a presentation of the system and they are suggested to use in their own free time as additional support. Its application, therefore, provided

[1]While it is understood that students' marks in assignments or exams and their knowledge cannot always be equated, students' performance, particularly from an invigilated exam and for particular questions, can be considered a good indicator of knowledge. For the purposes of the current research this was adequate as their previous knowledge was mostly taken into account as an additional factor during statistical analysis and data mining (e.g., to partial out the effects of previous knowledge in multiple regression).

an opportunity for observing how it is used and as a pilot for establishing its usability.

- `GIC02`

  By 2002 several small pilot studies had established the usability of the system and its acceptance by students for the `MM/AM` courses. Therefore it was considered possible to employ the system to facilitate the research described here. After developing some content particularly for the `GIC` course, it was used to supplement the teaching of GIC. As far as this research is concerned, this data-set was considered as a pilot in order to (a) establish the appropriateness of the material built and (b) collect data on unknown misconceptions in order to fine tune the feedback the system provides. In terms of the WALLIS project overall, this pilot was seen as an opportunity to investigate the potential of employing the system for summative assessments (Mavrikis and Maciocia, 2003a). 109 students out of the 117 who attended the course interacted with the system.

- `GIC03`

  Given the success of the pilot and after resolving some of the problems identified, WALLIS was employed more formally as the sole means of teaching conic sections. The online summative assessment was removed as it was shown to affect students' behaviour (this is discussed in detail in Section 4.3). Instead, students had to complete an assessment right after their interaction with WALLIS. Their mark in this assessment together with their mark in an appropriate question in the final exam were averaged and used as an indication for learning. As this was quite uncontrolled, it was only used to inform the machine learning investigations (described in Section 5.3). The data collected from students' interactions with the system were employed for the exploratory analysis of how students use the system presented in detail in this chapter. 126 students out of the 153 who attended the course interacted with WALLIS.

- `GIC04`

  Based on some of the observations from `GIC03`, the system was further fine tuned and a more formal study was possible to evaluate learning gains. In the following sections the terms learning and performance are used interchangeably and are quantified by averaging the marks students' achieved in an assessment they had to complete right after their interaction with WALLIS, as well as their mark on the final exam. Although the latter assesses long-term retention, it was considered sufficiently reliable for the purposes of the research reported here. The results may not be as reliable as when performing a more controlled experiment but are basically used as indications of the effects of students behaviours in learning. In addition, the marker of the exam was the author of this thesis and therefore learning attributed to the interaction with the system could be evaluated. The fact that the skill of converting a conic section to its standard form was covered using a particular methodology that students are invited to follow, facilitated the marking process. 133 students interacted with the system out of the 165 who attended the course.

- `GIC05`

  By 2005, based on the results of the analysis presented in this chapter, some of the functionalities of the system were redesigned (as Chapter 5 describes). However, the lecturer who was teaching the course was different than the previous years and he did not feel comfortable with using the system as the sole means of delivering the material. Therefore, this study was considered only a pilot to establish the usability of the redesigned system. Students marks in their assignment and their final exam were again used only to inform the results of machine learning analysis. 115 students used the system out of the 208 who attended the course.

Due to the way the data-sets were collected, some data are quite noisy and in some cases unusable. The method used to collect the data is subject to bandwidth availability, appropriate security settings and other server side concerns (see Appendix E). Therefore, they were preprocessed and in some cases the whole interaction of several students had to be excluded from the data analysis. In addition, some students did not give their consent for recording their data. Apart from these technical concerns there are other reasons to ignore data. For example, Section 4.3.1 explains that due to the lack of familiarity with the system data from students who did not attend the system's demonstration were ignored. In addition, the interaction of students who have taken the course in the past and failed are also ignored. After this data cleaning process the `GIC03`, `GIC04`, and `GIC05` contain 106, 126 and 99 students respectively.

4.3 Factors influencing students' behaviours

In section 2.4.3 several factors that influence students' behaviours, and particularly help seeking, were mentioned. While it is not always possible to control all of these factors, being aware of them played an important role in the methodology and context of the studies and analyses reported in this thesis. The next three sections describe the factors which, with some effort, was possible to control and therefore increase the validity of the results from the studies: *familiarity* with the system, the *feedback* that it provides and *task orientation*. Section 4.3.3 in particular contributes to a better understanding of the influence of task orientation by analysing results from data collected from early use of the system. As Aleven et al. (2003) describes there have not been many studies in the context of ILEs in relation to this factor.

Another frequently mentioned, and rather obvious but important, factor is *previous knowledge*. Since an educational system has to cope with different students, the factor should not be controlled (in the sense of excluding some students) but should definitely be taken into account when analysing the data. This will become evident in the subsequent analyses where prior knowledge is always partialled out to remove its influence on the reported results.

4.3.1 Familiarity with the system

Section 2.4.3.4 highlights the effect of students' familiarity with the system on their behaviour. A link between familiarity with the system and how it affects students' help seeking was identified at early stages of this research. In particular, it was noted that novice users of the environment who do not read the help pages appreciate the affordances of the system and the help that it provides with trial and error. This has lead the School of Mathematics to produce a special set of notes about the system and its use which are included with the students' notes when starting the Mathematical Methods course. This improved the way students use the system and particularly the help facilities.

However, the implication in the context of this research was that special attention was required so that students who participate in the studies familiarise themselves with the system in advance. A presentation of the system was given to them before every study and emphasis was given to all of the functionalities, and particularly the help facilities. In addition, in all of the studies all data recorded from the interaction of absent students were discarded. As also discussed in Section 2.4.3.4, it was clear that, in some cases, students were just clicking buttons and changing pages just as a way to explore the environment before starting to learn from it. The system's suggestions to seek help, students' curiosity as well as the information they got from the presentation affected their interaction. Therefore, all data analyses performed in this thesis ignore the first few pages students interact with.

4.3.2 Feedback provision

Section 2.4.3.4 identifies the importance of the exact feedback provided and how it determines students' help-seeking behaviour. Therefore, particular attention was given to the feedback delivered by the system.

While it is normal that not all hints or solutions are immediately understood by all students, an attempt was made to make sure that the feedback is clear for the majority of them. The pilot studies in the context of the WALLIS project, prior to this research, helped to fine-tune the feedback mechanism. More importantly, based on the `GIC02` and `GIC03` studies an effort was made, with the help of the lecturer of the GIC course, to ensure that most student following the pre-authored feedback provided would be able to complete the activities.

This observation about the influence of the exact feedback provided, had an implication in the data analysis performed, particularly the clustering presented in Section 4.4) or the analysis of various help-seeking measures presented in Section 4.5.3.2). Since the various feedback messages, hints and solutions are quite different from one another (in size, semantic density, difficulty of understanding them, etc.) any comparisons that involve feedback were drawn across these messages and then across each student rather than the more usual method of aggregating results per student and then comparing them. This is because, depending on their exact misconceptions and general behaviour, students may receive different hints rendering any comparison biased.

4.3.3 Task-orientation

As discussed in Section 2.4.3.3, another factor that seems to play an important role in students' behaviours is what is often referred to as task orientation. This is related to the teacher implicit or explicit messages about the goals of an educational situation. In the context of ILEs, Aleven et al. (2003) provide examples that demonstrate how students' personal orientation is manifested through their help-seeking behaviour. However, the influence of learning versus performance orientation should not be investigated simply as a student characteristic but has to be interpreted in the context of what the rest of the environment promotes.

In the context of WALLIS, observations of the way students interacted during the pilot application of the system during `GIC02` indicated that students are using the system generally in a less 'desirable' way than what was noticed during the initial informal observations in `MM/AM`. This less 'desirable' way can be summarised, using Nelson's (1985) term, as executive. As discussed in Section 2.4.1 executive help-seeking refers to those instances in which the students' intention is to have someone else solve the problem on their behalf. Although it was not possible to quantify these results, it was hypothesised that the assessment element associated with WALLIS in `GIC02` led students to a more performance-oriented interaction. Therefore, the assessment was removed.

Although investigating this issue in detail was out of scope for the current research and would require a more controlled experiment, a comparison was performed between the various data-sets. Due to the fact that these systems have some subtle differences only few of the actions are comparable. Therefore the comparison focuses on aspects of students' behaviour that according to Newman (1994) can be characterised as executive. In the context of WALLIS the following type of actions are considered executive-oriented:

- asking for help or the answer without first attempting to provide an answer,
- attempting to provide an answer just once and giving up the problem by immediately asking for the solution.

Based on this definition, the following ratio is calculated for each student interacting with WALLIS.

$$r = \frac{\text{total number of executive-oriented actions}}{\text{total number of all actions}}$$

To compare the data-sets an ANOVA analysis was conducted. An early investigation with GIC02 and GIC03 justified the removal of the assessment. The average ratios were $r_{GIC02} = 0.385$ and $r_{GIC03} = 0.304$ respectively. Although there is only a marginal statistical significant difference between the two groups ($F(1,153) = 3.99$, $p = 0.048$), it indicates that task orientation could be responsible for the different behaviour of students overall. After collecting all data-sets, an ANOVA analysis between all groups seems to strengthen this finding. The means of executive-oriented ratio for the GIC04, GIC05 groups were 0.356, and 0.664 respectively. Their means differ significantly $F(3,319) = 17.98$, $p < 0.05$ and a post-hoc Tukey$_b$ test, which is quite 'conservative' (Howell, 1990), showed that, with alpha at 0.05, only GIC03, GIC04, where task-orientation is the same, formed homogeneous subsets. Because of the uncontrolled nature of the study (different lecturers, no control for prior knowledge, and changes in the material) the result can be considered only as indication. However, the significant differences raise important questions for future studies.

Another aspect of students' interaction that seems to be affected is the order in which they access the pages. This behaviour is discussed in more detail in Section 4.5.1.2. Here it is worth mentioning that in both GIC02 and GIC05, students choose pages very strategically (i.e. choose first the assessment and use it as a guide on which particular parts to study). This is not surprising given what was discussed in Section 2.4.3.3 about the relationship between context, task and goal orientation.

The implication of the above is an important methodological point. Research related to students' actions can be easily compromised by such details such as the classrooms' general ambiance and the teachers' influences. These subtle differences can have an effect on any study and influence the results that are often reported between researchers. This also justifies the initial decision of being extremely careful of the context of the studies in this thesis as discussed in the Introduction.

4.4 Grouping students

The goal behind this part of the research was to determine groups of students that behave similarly. Based on these groups representative patterns of usage could be derived. The main rationale behind this was that deriving a few groups would allow the design of appropriate interventions for each group. In addition, this would reduce the complexity of the exploratory part of the analysis. Because many of the variables are interacting with each other, creating groups of students who behave similarly should facilitate the investigation in detail.

However, the size of the data and the many aspects that characterise students' interactions with the system made a manual grouping quite impossible and prone to bias. Employing data mining techniques such as clustering is a common approach for this task and is lately gaining ground in the field of AIEd (see some recent applications in Heiner et al., 2007, 2006). The task of grouping students according to their interactions with the system is similar to the task Chen (2000) faced in the field of web-based information systems where groups of users in terms of their interaction were formed successfully. Inspired by this work, variables (or *learner-system interaction* or *process* measures as they are called in Wood and Wood, 1999) were used to characterise a student's session. These are listed in Table 4.1. This list of variables is representative as it contains either directly, or indirectly, all the important aspects of a student's interaction with the system.

Their values were constructed automatically from the raw log files of the `GIC03` data-set by a log analyser (see Appendix E). However, some pre-processing was required. Apart from the general pre-processing mentioned in Section 4.2.3, some outliers had to be removed as they have the potential to seriously influence the reliability of the clustering. In particular, there were some students who stayed in an a page an implausibly long time compared to all the rest students. This was often accompanied

- Elements related to page-level interaction

1. A nominal variable representing the group in which students belong in terms of the order in which they accessed the material (see Section 4.4.1.1)
2. The Mahalanobis distance (see Section 3.4.6.1) of the time the student spent on each page from the minimum time of all students (i.e., the greater this value is, the greater the distance from the minimum time).
3. Page 'abandonment' frequency. The number of pages abandoned without completing their goal, over the total pages the student interacted with (pages that are abandoned immediately are not taken into account - see Section 4.5.1.3).

- Elements related to item-level interaction and help-seeking related behaviour.

4. A number indicating overall help frequency
5. A number indicating the time the student leaves between asking for hints and the previous event compared to other students (using Mahalanobis distance for vector) for the same hints.
6. The number of solution requests over solution exercise steps (if an mcq is exhausted it is considered as a solution request)
7. The number of theoretical material lookups that the student followed when such lookups were suggested by the system (-1 if no lookups were suggested)
8. The minimum estimated time spent reflecting after a hint (using Mahalanobis distance again for each hint). This time is only an estimation and is calculated from the time the hint was requested until the next action (e.g., next hint request, attempt to leave page or mouse movement outside the feedback area)
9. the tendency to ask for help rather than risk an error $\frac{help}{errors+help}$
10. The efficiency of solving exercises defined as $\frac{successful\_operations}{success+help\_requests+errors}$ Similarly to (Wood and Wood, 1999) in WaLLiS the learner is supported in solving any exercise so the number of completed exercises would not make sense, thus efficiency is a better measure to correlate with others. However, some students did not complete exercises or the goals of the pages. This is captured by the abandonment frequency.

Table 4.1: Session characteristics.

by a logging off action indicating that they had probably left the window open working on something else and then closed their browser, hence logging off. In some other cases where they seemed to restart their interaction after a long time, it was assumed that they did something different in between and therefore their actual time was replaced with the average time of all students. For the same reasons data from accessing a page and leaving implausibly quickly were also ignored. However, when removing outliers or ignoring data in favour of a statistical technique, it is worth taking into account any information that the may carry with them. In the qualitative analysis that follows (e.g., Section 4.5.1.1) all data were considered.

Another layer of pre-processing, that was necessary, involved some characteristics of the session (for example the order students accessed pages or the time spent on each page) which are represented as vectors over the whole student-system interaction. Including them raw in the clustering algorithm would increase the complexity of coming up with meaningful clusters. The semantics behind them require that they are treated as one characteristic but any clustering algorithm would use their individual elements as separate characteristics. This pre-processing is described in the next section. The rest of the variables, which encode general characteristics of the student (e.g., the number of requested hints over possible hints), are averaged across the items that the student accesses.

4.4.1 Pre-processing

4.4.1.1 Grouping the '*order of pages*' characteristic

The order in which students access the pages is an important characteristic because the skills that are involved in interacting with some of the pages are different. Accessing, for example, an exercise page first instead of the theoretical ones provides indications of the student's preferred learning manner. In addition, the kind of help one needs, the possible requests they can make and the number of reference pages they access during an exercise all depend on having accessed previous pages or not. Therefore, it was hypothesised that splitting the students first based on the order in which they access the pages, would help any patterns that are associated with this to emerge better.

Students who belong to the same cluster will have similar features and differences will be highlighted. To achieve this the following process was followed:

Data input

As Appendix B shows the GIC content for WALLIS consists of 13 pages. Consequently, 13-dimensional vectors were constructed containing as elements the i-th page the student chose for the first time and on which they remained for an amount of time substantial enough to interact or read what it includes. If the page was abandoned too soon[2] Weka's Cobweb algorithm is employed as discussed in Section 3.4.2. The outcome is three major clusters. The first two are further split into two clusters each.

1. Students linearly accessing the items as suggested by the suggestion mechanism or the order implied by the tree (59.43%). The two splits are those students who accessed all pages (Cluster **A1** 54.72%) and those who omitted the last few pages (which were more difficult and of less interest to the students because of the assessment they had to complete) (Cluster **A2** 4.71%).

2. Accessed the theory first, then the exercises and then examples (25.47%) Again the split was on accessing the last additional pages (Cluster **B1** 22.64%) or not (Cluster **B2** 2.83%).

3. Accessed first the exercise or the example and then sometimes the theory. Some accessed some of the last material particularly the exercises (Cluster **C** 15.09%).

4.4.1.2 Per page characteristics

Some of the characteristics that were included in the clustering process relate to aspects that should be compared across students and across items they access. One simple approach which is often employed is to average this aspect of the interaction (e.g., average time per page) but this seemed problematic since the items differ a lot and only one-to-one comparison makes sense. The same applies for other characteristics: for example, the time between hints. Averaging across all hints a student requested

[2] *Too soon* is defined as $t < -1.28$ standard deviations below the mean time of leaving a page. Section 3.4.6.2 describes the discretisation process and the rationale behind pageordering

would not offer a fair comparison with the average of another student who requested different hints.

In such cases one approach is to estimate the standard deviation of the distances of the points under question from a mean or centroid value. However, this assumes that the points are distributed normally around the centre of mass. This is not the case for all the variables here. Therefore, in order to have more realistic comparisons, the Mahalanobis distance was used as described in Section 3.4.6.1 was used.

Subsequently, the data were pre-processed in the following manner :

1. For every characteristic and for each student, a vector was constructed. For example, for the variable 'time spent on each item' a vector of 21 elements was constructed (the `GIC` content consists of 21 separate items) with the time the student spent in each element.

2. A vector with the minimum values of each element from the aforementioned vector was constructed.

3. The covariance matrix of all the vectors of all students was calculated.

4. The distance between each vector and the minimum vector was calculated. This value was used as the student's characteristic during the clustering.

After performing the pre-processing, the numerical variables were further discretized (based on the technique described in Section 3.4.6.2) to facilitate the hierarchical clustering which handles nominal values better (Witten and Frank, 2005).

4.4.2 Results

The clustering process yields 18 different groups of students. The representative vector and the size of each cluster is presented at table 4.2. The large number of clusters is contrary to the original expectations. Although further grouping is possible, the variance of some variables inside some of the clusters (either small or large) is already too large to be able to understand the distinguishing characteristics of each group. This makes the task of designing interventions for students belonging in each of the clusters difficult. This also demonstrates the complexity of students' interactions and that a more thorough investigation was needed. The patterns emerging indicate that further statistical analysis was needed to elucidate important aspects of students' behaviours and the interaction between the variables used to characterise them.

However, the results were helpful in seeing emerging patterns and identifying students that was worth interviewing or just looking further into aspects of their interactions. These interviews played an important role in the results mentioned in the next section. Choosing randomly which students to select would have been difficult and would require a large sample. Having performed the clustering (despite the large unexpected number of clusters) helped to identify outliers and centroids worth investigating further.

Cluster/Variable	*Size*	*page order*	*time*	*page abandonment*	*help frequency*	*between hints*	*solution requests*	*material lookups*	*reflection*	*help tendency*	*solution effectiveness*
Cluster 1	11	*	M-VH	*	ML,HL	M-VH	M,H	M,H	VL-M	M-H	L,M
Cluster 2	13	*	M-VH	*	ML,HL	M-VH	M,H	L	VL-M	M-H	L,M
Cluster 3	4	*	M-VH	*	ML,HL	M-VH	L	*	H-VH	L-M	H
Cluster 4	7	*	VL-L	*	ML,HL	M	M,H	*	*	L-M	L,M
Cluster 5	3	A1,B1,B2	VL-L	M	ML,HL	M	L	M	M-VH	M	M,H
Cluster 6	3	A1,B1,B2	VL-L	M	ML,HL	M	L	M	VL-M	M	M,H
Cluster 7	3	A1,A2,B1	M-VH	M	LL	M,H	M,H	H	M	M	L,M
Cluster 8	3	A1,A2,B1	M-VH	M	LL	M,H	M,H	H	M	L-M	M,H
Cluster 9	2	A1,B1,B2	VL-L	M	LL	M	M,H	H	L-M	M	L,M
Cluster 10	5	*	M-VH	L-M	LH	M-VH	*	*	M-VH	*	*
Cluster 11	7	*	VL-M	*	LH	VL-M	*	*	VL-M	*	*
Cluster 12	8	*	M-VH	*	MH	VL-M	M-H	VL-M	VL-L	M-H	VL-M
Cluster 13	14	*	VL-M	*	MH	VL-M	M-H	VL-M	L-M	M-H	VL-M
Cluster 14	17	*	VL-M	*	MH	VL-M	M-H	VL-M	M-H	M-H	VL-M
Cluster 15	7	*	VL-M	*	HH	M	M-H	VL-M	VL-L	M-H	VL-M
Cluster 16	3	A2,B1,B2	M	M	HH	M	M-VH	L-M	VL-M	H	L
Cluster 17	6	*	VL-L	*	HH	L,M	M-VH	L-M	VL-M	H	L
Cluster 18	3	A2,B1,B2	VL-L	M	HH	L,M	M-VH	L-M	VL-M	H	L

Table 4.2: Vectors of the clusters obtained from COBWEB with possible values of the session characteristics. Star (*) indicates that the variable takes any value in its domain, otherwise the specific range of values for every variable are shown. For example, for the `page order` variable the possible ordering of pages is indicated (see page 73). For the other variables, `L,M` indicates that the variable takes only `low` or `medium` values, `M-VH` indicates that values from `medium` to `very high` are possible.

4.5 Students' behaviours and their relation to learning

The following sections present various aspects of students' behaviours with the system by employing the observations from replays of students' interactions in GIC03 as well as the interviews performed as a result of the clustering presented in the previous section. Based on these results, 3 students from each cluster were asked to attend a semi-structured interview. Only 25 of them attended but this provided sufficient representatives from each cluster to give enough variability. During the interviews students were shown a replay of interesting parts of their interaction and were asked to comment on any aspect they found difficult, and provide feedback on the content and help they received from the system. The interviews were quite open, leaving enough freedom to students to raise any issues. However, having watched their interaction in advance and knowing in which cluster they belong helped identifying key questions for each one of them in order to find the underlying reasons of some of their actions, especially the actions which were not anticipated while designing the system. These interviews helped identifying the issues that are discussed in this section. Whenever possible, statistical techniques are employed to support the qualitative findings. Although the analysis is mostly focused on help seeking behaviour other types of interactions (such as page navigation and response giving) were also analysed. In the sections that follow their relation to students' performance is described and, wherever possible and useful, a comparison is performed with results from similar research.

4.5.1 Navigation

Navigation is a feature that distinctly differentiates most web-based ILEs from traditional ITS. The fact that students are free to navigate through the material and choose their own learning section from what is often a broad coverage of material provides them with more control during their learning experience. However, as Thuring et al. (1995) and others (e.g., Brusilovsky, 1996; Conklin, 1987) also highlight, students are often lost in the choices and are not sure what to do. Many adaptive learning environments target this problem by adapting the available links to the students, hiding already visited pages or other pages which the system assumes that the students would find hard to interact with, or expanding links with definitions (Brusilovsky, 1996; De Bra and Calvi, 1998; Eklund, 1995). WALLIS, as described in Section 3.2.1, employs a similar approach where links are annotated with an icon that is empty, partly or totally filled when students have interacted to indicate the completion of the item's goals.

However, as is often the case in many eLearning environments, the fact that users have more control over the material they choose, makes them develop their own way of interacting regardless of the intentions of the system designer. Students often access pages in a different order, using the back button of the browser to visit previously visited pages. This way, any solution adopted, although helpful for the students, does not necessarily resolve the problem of the false evidence that the system takes into account when modelling them.

The following sections describe further the way students select pages, the order they choose and quit them as well as the relation of these aspects with learning. Wherever possible issues are raised that should be addressed in the system's redesign.

4.5.1.1 Selecting pages

A pattern emerging from the data analysis and the replays of student interactions is that some students wander from page to page apparently without a specific task in mind. In all GIC data-sets an average of 31.89% finished the interaction with the system with having at least one page read too quickly to have possibly understood its contents carefully enough or abandoned at least one interactive exercise before completing its goal. As will be described in more detail in Section 4.5.1.3, a large number of students abandon pages the first time they interact with them. Initially, it was considered that this was a wrong choice that students made and that as soon as they realised what the page contains they abandoned it to go to the one they wanted to select. The large number of instances though and the fact that the same behaviour is repeated over different years, does not justify that. The replays and interviews helped to identify that, apart from wrong choices, there are other reasons behind this behaviour.

Students do not always know in advance what to expect from a page and how it relates to their current goal. Since they are not sure of the reason behind interacting with it, they often access a page just to see what it involves, they quickly scroll or start interacting with the items but then leave to go back to the appropriate theoretical page or example.

4.5.1.2 Order of pages

While authors of eLearning content usually design stand alone content, prerequisites are unavoidably assumed. Depending on the context, students come with some background knowledge and the author usually takes this into account. Consequently, this has an effect for the system in terms of modelling students' knowledge. Students do not always interact with all the available material in the order expected. For example, in situations where WALLIS is used just as support material (like in the MM data-set) an average of 64.52% of the students, immediately access examples or exercises before looking at the relevant theoretical pages. This is because they have usually acquired some knowledge in the classroom and they prefer to interact with the examples. A traditional ITS (as discussed in 2.1) would not necessarily take this into account and therefore the system would try to force the students to interact with the theoretical material first to cover any gaps that they may have. However, even in situations where the material were not taught in class (like the GIC data-sets) and students were not aware of them in advance, as the results at Section 4.4.1.1 also indicate, there are still some students who interact first with exercises or examples rather than theoretical pages. In all GIC data-sets this was on average 40.32% of the students.

In addition, during the interviews many students who follow this approach commented that it helps them to get an overall feeling of what they are going to work with. It became evident that this is an approach to learning that they have in general. More specifically, for some of them, this was not a random decision but an explicit choice as they said that they were used to have this approach(from school) of skimming through material to get an overview before focusing on the individual sections. This behaviour is not necessarily problematic but is rather related to students' learning styles. For example, the two most widely used inventories of learning styles of Kolb (1984) and of Honey and Mumford (1986) both recognise students' preference for starting with examples first rather than theory. Also in the classification of learning styles of Solomon and Felder (1998) and Felder and Silverman (1988) this behaviour

		Cluster					Total
		A1	A2	B1	B2	C	
	N	65	11	25	8	17	126
	mean	5.557	5.143	5.048	5	5.385	5.632
	std	2.029	1.952	1.774	2.16	1.85	2.184
t-tests for prior	t(124)	.408	.611	1.373	.543	.434	
knowledge	sig	.684	.542	.172	.588	.665	
multiple regression for	r(123)	.108	-.022	-.003	.025	.032	
post-test performance	sig	.231	.808	.974	.782	.723	

Table 4.3: t-tests of difference for prior knowledge among the different navigation clusters (established in Section 4.4.1.1) and multiple regressions against post-test performance partialling out prior knowledge. No t-test or correlation is significant.

differentiates between *Sequential* and *Global* learners. Global learners are more holistic thinkers who require larger steps and are good at synthesising the different parts. It seems that students have transferred this approach to the computer based environment. Similar behaviours have been reported elsewhere (Coombs, 2006; Joyes, 2006; Wong et al., 2007).

However, this behaviour is not necessarily problematic. First of all, it seems to be a behaviour that does not depend on students' previous knowledge. In GIC04 there was no significant difference between means for prior knowledge and belonging to one of the clusters in terms of order of accessing pages (see Table 4.3). In addition, the cluster that a student belongs seems to have no significant impact on performance. All multiple regressions performed in order to partial out previous knowledge, result in non-significant and very small correlations (see Table 4.3). This substantiates the claim that students should be allowed to access the pages the way they prefer and that the system could just play a supportive role.

4.5.1.3 Page Abandonment

As described in Mavrikis and Maciocia (2002, 2003b) in very early versions of the system a high proportion (56% on average) of students did not complete all goals. Although the completion of activities is one of the easiest aspects of student interaction for the system to monitor, the exact reasons for abandoning pages are different for every student and therefore need to be interpreted carefully.

The fact that students quit a page does not necessarily imply something in relation to their motivation or the page itself. In some systems this is interpreted as such (for example, Andres et al., 2006; Aist et al., 2002).

First of all, as already mentioned in the previous section, a careful analysis of the log files in conjunction with interviews revealed that in most cases students just want to see what certain pages involve (i.e., if it is a theoretical page or something they can interact with) and then immediately leave (for example, to come back later). A large number of pages are left so quickly that it would be impossible for students to have understood details about the page but rather just get a general feeling. For example 9.4% of these cases are pages that are abandoned too soon. This is too high a proportion of students to ignore.

Other students abandoned pages immediately after seeing a step of the question and either attempted to answer wrongly or realised that they could not answer. Despite the difficulties they were facing and the fact that they may have already studied the appropriate pages, they did not ask for more help. The reasons behind this type of abandonment are different and therefore different actions are required from the system.

Recognising the complexity of this issue and that is not easy to resolve a first approach (during `GIC03`) was to add a simple mechanism that pops up a prompt that proposes to the student to remain on the page and ask for more help. This simple mechanism served the purpose of making students reconsider abandoning the page and focused their attention more explicitly on this process. Apart from reducing the abandonment rate to 14%, this small intervention helped in getting more results during the interviews.

From the interviews it was determined that students abandoned pages mostly because of the following:

1. the goal of the activity and its relevance to their learning process was not obvious to them, or

2. they were successful in early parts of the activity and felt really comfortable with the rest, or

3. because they were dissatisfied from struggling without managing any of the goals and wanted (a) to try to finish some other time, or (b) visit the appropriate page where theory is covered and (c) come back later or, having lost interest and motivation on this page, never came back.

These reasons are all related to affective and motivational characteristics and particularly confusion, boredom and lack of specific goals. The interviews also raised important interface issues such as the fact that certain students (especially the ones visiting a page in order to see what it contains, or the ones leaving to come back later) were annoyed by the prompt. When told that this could improve the system's ability to suggest study material and adapt the feedback they were still quite concerned by the interruption that would occur but some said they would not mind if this was reminding them and helping them to learn more from the page. Similar results are reported in de Vicente (2003) where students preferred self-updating their motivational model at the end of each interaction rather than in the middle.

The `GIC04` data show a very small negative correlation ($r(124) = -.029$) between learning and explicit page abandonment (i.e., abandoning the page even after the prompt mechanism). This correlation is not significant. However, as one would expect, performance in skills of which the corresponding sections were abandoned in this manner is significantly lower. This is better demonstrated when controlling for prior knowledge ($r(502) = -0.15, sig < .05$)[3]

4.5.1.4 Implications

In traditional ITS the more structured sequence of material helps prevent several of the problems reported above. However, too much control can hinder the sense of locus of

[3]When correlating for individual skills the following approach is taken: There are four skills (1) recognising equation (2) finding matrix, (3) finding the standard form (4) diagonal, and the various variables are calculated across these skills

control which, as mentioned in Section 2.3, in particular adult learners usually expect to have. The advantage of ILEs that foster more learner control could be strengthened by enhancing them with the ability to provide, apart from simple suggestions on what the student should study, explanations of the reasons behind the suggestions and how the selected page would contribute to the student's learning. Chapter 5 describes how the suggestions of the system were changed to accommodate these problems.

Related to the problems occurring from not allowing students to select material, not allowing them to choose the order they want to read material could also lead to frustration or boredom. It is important to set up the system in such a way that it will not influence unnecessarily both their learning and affect. On the other hand, allowing them to access the material on their own increases the self-regulated skills and allows them to explore the material in their preferred way. This sense of self-directed learning is related not only to satisfaction but also to engagement, curiosity and hence increased motivation (Lepper and Woolverton, 2001; Arnone and Grabowski, 1992).

In addition, structuring a specific order is not necessarily beneficial. Therefore, students could access the pages the way they prefer but the system should not necessarily consider just looking at a page as evidence of having learnt the material there. Just looking over a page or even spending some time with it does not necessarily provide evidence of having learnt the appropriate material. A different mechanism is needed that could allow students to interact first with their preferred material and even skim through the pages, and evidence on a page being abandoned should be treated with low confidence. This behaviour mostly points towards student characteristics (e.g., their learning style) rather than anything else. Once misconceptions are identified, the suggestion mechanism could divert students to pages that they have not covered adequately.

Finally, it was identified that the reasons behind abandoning pages are all also related to affective and motivational characteristics and particularly confusion, boredom and lack of specific goals. When taken into account they could facilitate the system's ability to diagnose and target affective aspects. Given the acceptance and effectiveness of the simple prompt mechanism described above, Chapter 5 describes the improvement of the prompt that not only makes students think more explicitly about their actions but also offers a practical improvement to the interface. In addition, Chapter 6 shows how the answers to the prompt contribute to the system's ability to diagnose affective characteristics and particularly *effort*.

4.5.2 Response giving

The responses and the way students provide them also constitute useful evidence to be taken into account by the system. Apart from evidence in terms of cognitive aspects (e.g., misconceptions) the exact interaction can facilitate the affective diagnosis process. The most interesting aspect of a student's way of inputting answers into the system is the evident hesitation of some of them while answering. Tutors in the human-student study described in Section 2.5 expressed their desire to intervene in cases where the student was hesitating (Porayska-Pomsta et al., 2008). Intervention to boost students' confidence to attempt to answer as a tutoring approach is very useful and often suggested to student tutors (e.g. Forster et al., 1995). Further to the exact way of answering, the following sections describe different types of events in relation to the answer itself.

4.5.2.1 Random answer

Observation of the answers the students provided, together with the interviews, show that some students answer questions randomly. This is obviously not desirable interaction, particularly because of the fact that in most of the cases the system cannot deal with the error and therefore cannot help the student at all. Having the ability to detect these random answers would be very useful. However, this is not so straightforward as it is hard to differentiate between random answers and some complex misconceptions that a certain student may have. More detailed analysis is needed to pinpoint which answers are indeed random in order to use this aspect as a means of designing interventions. However, a good predictor of random answers would require detailed and focused research on its own right. For example, Beck (2005, 2004); Mostow et al. (2002) have dedicated a lot of effort investigating this and similar aspects of the interaction in detail in the context of a reading tutor. Since this is not the main concern of this thesis only a preliminary approach was chosen to investigate the possibilities that predicting random answers could have. A difficult skill, that of finding the eigenvalues, was chosen. The skill requires a substantial amount of time to answer, even if someone knows the method for answering it. In addition, most of the possible misconceptions are well known due to the consecutive years of use of this particular question. Therefore, whether an answer is random can be determined by hand after removing all answers that are known misconceptions and taking into account the time between the previous event and the actual submission of the answer measured in standard deviations from other students. While there are probably random answers given after some time has passed and some answers that by luck look like common misconceptions the manual method described provides an approximation to investigate the issue further.

Based on this definition, on average (over all data-sets) only 4% of students who's first attempt seems random corrected their answer after receiving help. This indicates that they are answering randomly just as way to make the system believe that they are providing an answer. This allows them to request the solution or more hints. It is interesting to observe that these students tend to have lower previous knowledge ($N = 10$, $\mu = 3.4$, $\sigma = 1.776$). A t-test on these data is significant ($t(122) = 3.387$ and $sig < 0.05$).

In the `GIC04` data-set, belonging in the group of students who answer randomly correlates negatively with learning ($r(121) = -.177$); a correlation that is marginally significant ($t(121) = -1.984$, $sig = .05$). Restricting the set to include only students who seem to have answered randomly more than once increases the correlation to $r(121) = -.271$ and this time it is significant ($t(121) = -3.094$, $sig < 0.05$) [4]. These results, although not surprising, provide support for the claim that this behaviour should be targeted.

4.5.2.2 Copying from example

As noted before students were able to access examples as a means of help during the exercise. Replaying the interactions of some of them and during the interviews it was identified that answers are sometimes copied from the examples provided with the as-

[4] Due to the manual nature of this and the following analysis only one skill was tested for this correlation.

sumption that the answer transfers to the exercise the student is currently working with. This was particularly the case for the eigenvectors step (page SF-P in Appendix B) which does in fact look very similar to the relevant example. In total 8 students who tried this step seem to have copied the answer from the example. This is determined by considering if the answer is the same as the example and if the student has previously read or is currently reading the example page. These students, like the ones who answered randomly, seem to have low previous knowledge ($N = 8, \mu = 3.125, \sigma = 1.727$). A t-test indicates that this is statistically significant ($t(122) = 3.386$, $sig < 0.05$). The correlation of this aspect of the interaction with performance (partialling out previous knowledge) is negative $r(121) = -0.156$ but not significant ($t(121) = -1.742$ $sig = .084 > 0.05$). Excluding the 3 students who corrected their error, based on the feedback they received, and answered correctly immediately, the correlation increases $r(121) = -.260$ and becomes significant $t(121) = -2.957$ and $sig < 0.05$.

4.5.2.3 Speed of answering

The speed of answering the question after it was displayed or after a hint was given can provide evidence for not reading the question or the hint carefully. In addition, as established previously, it seems to be possible to identify the random answers based on the speed that they are given.

Although there is an indicative negative correlation $r(498) = -.086$ between the speed of answering the steps and performance (again controlling for previous knowledge) this is not significant $t(498) = -1.926$, $p = 0.055$. However, the interviews revealed that this aspect of the interaction is quite complicated as it depends on many other factors. For example, the difficulty of questions plays an important role. The same analysis for the most difficult question yields a negative correlation $r(124) = -.197$ which this time is significant $t = 2.238$, $sig = 0.027$. The interviews with the students revealed that other factors such as the knowledge students have just acquired by the interaction with the system play a role in the speed a student is answering the question. In section 4.5.3.2 the role of time is analysed further in relation to the speed of asking and reading hints as well as how it relates to learning.

4.5.2.4 Self-correction

There were two types of self-correction: one that occurred before even submitting the answer and one after feedback was received. Self-corrections are interesting because tutors use them as evidence of an increase of confidence (Porayska-Pomsta et al., 2008).

There was no significant correlation between students' performance and this aspect of their behaviour. The small sample of 12 students, who correct themselves only once or twice in average, probably hinders any patterns from emerging. However, what is evident is a difference in previous knowledge when comparing a group of students who self correct at least once ($\mu_A = 6.73$, $\sigma_A = 1.668$) with those who never self-correct ($\mu_A = 5.396$, $\sigma_A = 2.188$) $t(124) = -2.276$ and $p < 0.05$. Considering only the students who self correct and have low previous knowledge, they have an overall satisfactory performance $\mu = 60.94\%$ but the difference with the rest of the students is not statistically significant $t(55) = -.134$ $p > .05$. The effect therefore, could be just random. However, there seems to be scope in using this behavioural aspect for modelling purposes.

4.5.2.5 Implications

Chapter 5 will present in detail how the above aspects are taken into account. While random answers are difficult to predict, self-corrections and the speed of answering are explicitly taken into account in identifying student's beneficial interaction. In addition, despite the fact that copying from an example is not directly taken into account in the modelling process of beneficial interaction, it is directly linked with an intervention from the system (see Section 5.4.3.3).

4.5.3 Help-seeking and related behaviours

Chapter 2 established that very little empirical research exists on help-seeking in ILEs. The few results mentioned in Section 2.4.2 from (Wood and Wood, 1999; Schworm and Renkl, 2002; Aleven et al., 2003) are sometimes contradictory and not conclusive. The data from WALLIS provide the opportunity to investigate these aspects further and contribute to the overall understanding of which aspects of help-seeking are related to learning and how. Prior to the analysis performed here it was expected that some of the results would be replicated. As the rest of this section shows some of the results challenged previous findings. However, even some of the findings that replicate previous ones still a contribution since they are investigated in a different system, situation and context. The target group (first year university students) is quite different from most of the studies above. Therefore, by performing these comparisons one of the goals mentioned in Aleven et al. (2003) is targeted; that of determining to what extend do specific help-seeking actions generalise across different environments.

This is partially the goal of the next section that describes the general usage of help in the system and its relation to other findings from similar research. Section 4.5.3.2 focuses on several learner-system interaction measures (see page 64) and particularly on the frequency of help-seeking and its relation to learning, with the goal of assisting the subsequent model building process in the next chapter.

4.5.3.1 Help use in WALLIS

As discussed in section 2.4 when designing ILEs or ITS with feedback, designers expect that students request help from the system when they need it. However, the relevant research, mentioned also in 2.4, indicates problems that need to be addressed.

Moreover, from the clustering process, the replays and interviews it is evident that there are some groups of students which manifest a behaviour quite different from the one expected when designing the system. Students request hints without spending enough time on the step first, neither attempt to answer nor reflect on hints before requesting the next one. In addition, they ask for solutions too soon or seek for the most explicit hint that almost gives the answer a way . This is contrary to the evidence of hint-avoidance (i.e., students not requesting hints when needed) reported in Wood and Wood (1999). Aleven and Koedinger (2000) document some similar results, of students working with the CMU tutors. Although they also highlight that, in their data, there is a tendency, when requesting help, to exhaust all possible hints, they also report that when students made an error, and if they had not asked for help already, they were more likely to attempt another answer than to ask for help. This can be seen in Figure 4.1. For example, after making 2 errors without asking for help student asked for a hint only around 33% of the time. In the WALLIS data the results suggest that students ask for help more than they should. Figure 4.2 illustrates the difference. It shows the frequency of attempts to answer after *N* errors without any prior help request. For example, after making 2 errors only around 22% of the next step were attempts to answer. After 3 errors an impressively low percentage (around 3%) of next steps was an attempt to answer. There were no attempts to answer after 4 consecutive errors.

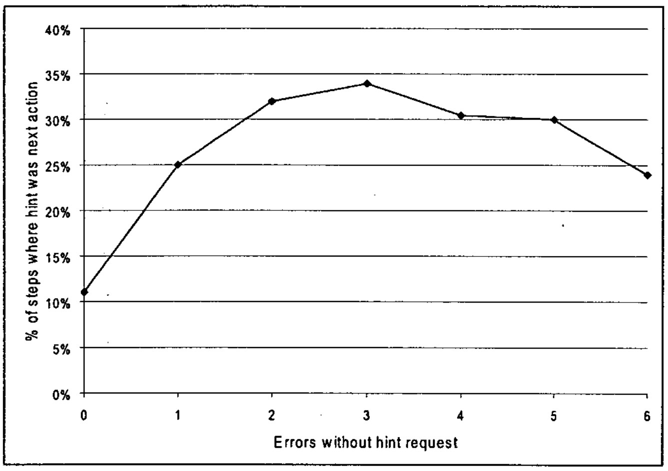


Figure 4.1: Frequency of help use as a function of number of errors without help request (adapted from Aleven and Koedinger, 2000)

In addition, contrary to Wood and Wood (1999) where students tend to avoid asking for help, in WALLIS students tended to abuse the hint feature. For example, only around 25% of the students attempted to answer when they were presented with a question, the rest requested a hint immediately. Similar patterns, of students' abusing help facilities are reported in Baker et al. (2004b) and Baker (2005). This behaviour is referred to as 'gaming the system'.

It is interesting to interpret the different results. The developmental or even cultural differences of the students could be playing a role. The undergraduate students are

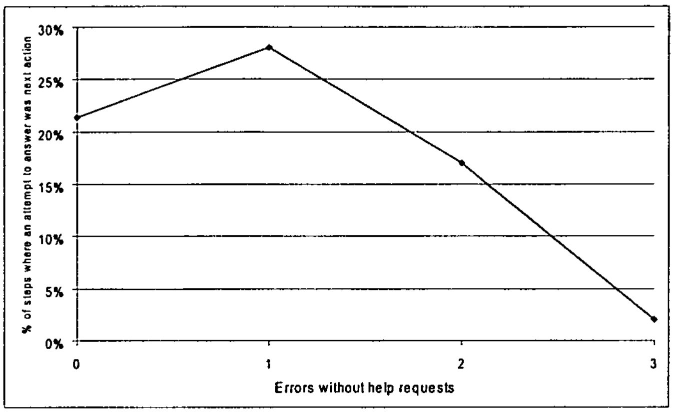


Figure 4.2: Hints after errors in GIC04

metacognitively more aware of their need for help than younger students (Schoenfeld, 1987). On the other hand, the results here indicate that they ask for more help than necessary. It is possible that this behaviour is an artifact of the system and the context. In CMU's cognitive tutors there is a penalty for asking hints in the student model and after a while perhaps students realise (or perhaps they were told) that their progression depends on the amount of hints they ask. In WALLIS not only there is no penalty associated with hint-requests but also the hint button is located closer to the text field where students are supposed to provide an answer, perhaps priming them to ask for more hints than they require.

Another possible explanation for the differences is the amount of time students had to use each system (Koedinger personal discussion, 2005). Indeed, the data collected in this thesis (and in Aleven and Koedinger, 2000) are from situations where students have already spent time working with the system. Therefore, the way they interact with it has been shaped as a result of their previous interaction. For instance, it takes two or three activities for the students to realise that exhausting the hints gives them the opportunity to request the solution. Another reason for the differences could be the fact that students in Wood's studies participated in a study that was set up in a lab. All the studies reported here, as well as those in Aleven and Koedinger (2000), were from situations where the system was integrated into the curriculum. This provides different goals for the students and as discussed elsewhere in this thesis (e.g., in Section 2.4.1) this is an important factor to consider when researching aspects of interaction that are related to motivation.

4.5.3.2 Help-seeking measures and their relation to learning

Frequency of help-seeking

A learner-system interaction measure that is very often discussed in literature is the frequency of help-seeking. However, controversial results are reported. Wood and Wood (1999) found that students with lower prior knowledge who seek help more frequently tend to have higher learning gains. On the other hand, Aleven and Koedinger (2000) report an overall negative correlation between help seeking and learning. In an attempt to perform a similar analysis an obstacle was met. The help seeking frequency of the data in WALLIS does not follow a normal distribution at all, thus violating the assumption needed for multiple regression. The exact reasons behind this are not immediately clear however by plotting the density of the help-seeking frequency (see Figure 4.3) it becomes apparent that the variable follows a bimodal distribution. Controlling for the obvious factors such as gender or prior knowledge does not seem to differentiate the two modes. More specifically, by splitting in the two modes based on the medium value ($\mu = 48.09$) a t-test ($t(502) = 0.590$) is not statistical significant indicating that prior knowledge cannot differentiate between the two modes.

Consequently, it seems that other characteristics of the student (not necessarily captured in the data set) play a role leading to this result. For example, since help depends a lot on metacognitive skills and since student with both 'low' and 'high' previous knowledge could have this skill, this could be the reason behind this bimodality. Other factors could be students' learning style or approach to learning in general. In addition, affective and motivational characteristics could be the underlying this behaviour. At this stage, before attempting to pinpoint the exact reasons behind this, it was more important to identify the relation of help-seeking frequency to learning.

In order to perform such an analysis, the data were split in two groups that follow normal distributions. Grouping the students with help frequency values below the mean amounts to 41.27% of all students. This data set comprises Group A and follows

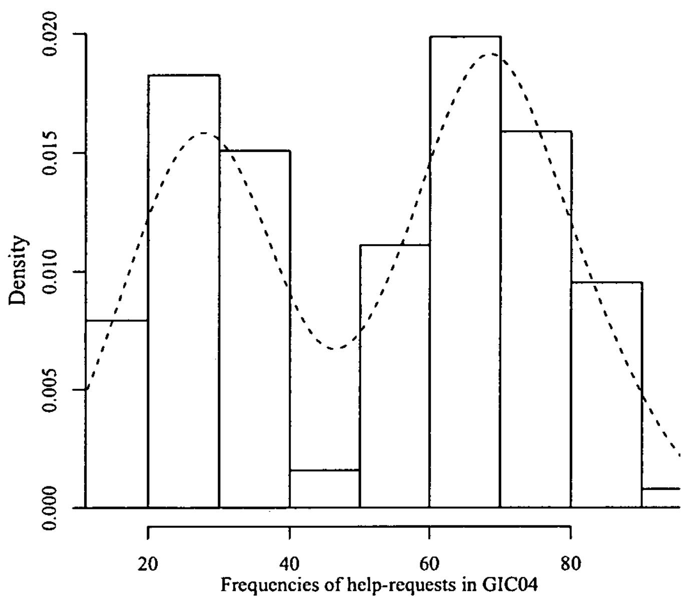


Figure 4.3: Density and histogram of help-seeking frequency.

a normal distribution with mean $\mu = 27.32\%$ and $sd = 6.967$. Group B includes the remaining 58.73% and also follows normal distribution with $\mu = 68.71\%$ and $sd = 10.105$.

Now it is possible to investigate how prior knowledge relates to help-seeking.

- For group A there is a negative correlation $r(206) = -.132$. However it is not significant ($p > 0.05$).
- For group B there is a negative correlation $r(294) = -.149$ which is significant ($sig = 0.01 < 0.05$).

This is in accordance with similar results (Wood and Wood, 1999). In general students with less previous knowledge ask for help more. This is consistent in both Wood's and Aleven's results.

In terms of learning:

- Help frequency for group A is positively correlated with learning (controlling for prior knowledge) $r(206) = .349$ and this correlation is statistically significant ($t = 7.374$, $p < 0.5$).

- In group B although help frequency is negatively correlated with learning (again controlling for previous knowledge) $r(292) = -.094$ this correlation is not statistically significant $t = -1.894$, $p > 0.59$.

The results demonstrate that there is a substantial amount of help seeking going on and a certain amount of it is beneficial under certain circumstances. However, some aspects of the group who seeks help with higher frequency (group B) seem negative in terms of their learning. In addition, in each group the frequency of help seeking has the potential to help in the prediction of performance. However, because of the fact that we would need some reliable way to differentiate between the two modes, help seeking on its own cannot be used for modelling beneficial interaction. Some other student characteristic is responsible for this separation. One hypothesis is that the behaviour is related to metacognitive abilities in general, learning styles or general orientation towards learning (e.g., performance-orientation). Unfortunately such student characteristics were not available at the time of this analysis. The interviews with the students were conducted before this aspect was fully investigated. However, as discussed in section 2.4.3.2, an attempt was made throughout this thesis not to include such information in the model as they would have to be collected using questionnaires and self-reports in advance. It is therefore considered more useful to be able to derive predictors of beneficial interaction based on the actual actions. This is investigated in the next chapter.

Another hypothesis, derived from the interviews, is that affective and motivational characteristics are behind this behaviour. Although, at this stage, this was not explicitly investigated, students often mentioned factors such as boredom, confusion, interest.

Bottom-out hint

Another behaviour that is often discussed in relation to help-seeking and ILEs is students' request for the last hint and for solutions. On first sight, this aspect of students interaction does not seem desirable. For example, it is one of the behaviours termed 'gaming the system' in Baker (2005). A closer look at the data however does not clearly demonstrate this. While there is a high negative correlation between belonging to this group and learning this is not statistically significant, nor does there seem to be any difference in prior knowledge. However, some of the students who request the bottom-out hint do seem to learn the associated skill. Looking into the detail of the data, it seems that this is related to the time students' spend reflecting on the hints. This is another aspect that was identified after the interviews were conducted and therefore it was not investigated in detail. Nevertheless, it seems plausible that some students treat the solutions very meticulously and do learn from them. It seems that looking in more detail at the time they spend on hints is more important than the frequency with which they request them.

Speed of asking for hints

Students in group B take shorter time to ask for the first hint than students in group A. Two ways of verifying this were employed. The more general comparison is to compare the Mahalanobis distances of the vector from the vector of minimum times. Group A has clearly ($\mu_A = 99.082$, $\sigma_A = 9.737$) larger Mahalanobis distances from Group B ($\mu_B = 81.6483$, $\sigma_B = 13.86004$). This difference is statistically significant ($t(124) = 7.812$, $p < .05$). In order to be more precise, sixty t-tests for every hint were performed. Apart from two, all tests were statistically significant with $t(124) > 1.98$. The two tests that were not statistically significant, were from the exercise with the most difficult steps in the system. This implies that in this difficult step even the students in Group A asked help more quickly.

One could hypothesise a similar solution to that used for the abandonment aspect of the interaction (i.e., a prompt suggesting not to ask help so soon) would be helpful in stopping students' behaviour. However, looking at the data above (1) this behaviour is not necessarily harmful and (2) it would be too much intervention.

Interviews with the students raised an issue that it is difficult to take into account in a simple statistical analysis as the above. The interaction with the system prior to interacting with a specific step or exercise plays a significant role. This was also mentioned in Section 4.5.2.3 in relation to the speed that students were answering questions. This should be expected. Due to the way the content is designed some students (regardless of their prior knowledge) learn from the examples or the theory before accessing the exercises. In addition, the interaction with the other skills, for example prerequisites or similar exercises, often help in their general understanding of the subject matter. Therefore, the knowledge that students acquire while they are working with the system needs to be taken into account before deciding if their request for help is superfluous or not. An attempt to take this factor into account is described in Section 5.3.1 that presents a model which predicts whether a help request seems necessary given their previous interactions.

Hint reflection

Students in Group A reflect more on hints. Performing a t-test on the Mahalanobis distances of the times students reflect on hints (as indicated by their actions after requesting a hint) shows that Group A reflect more on hints overall ($\mu = 1629.983$, $\sigma = 57.419$) than students in group B ($\mu = 1333.739$, $\sigma = 29.76$). This difference is also significant ($t(124) = 8.341, p < .05$). The same was true for all hints apart from the last one for the two difficult exercises.

4.5.3.3 Implications

From all of the above it is evident that overall students do use the help features of the system and it seems that, for some of them, the help they receive from the system is beneficial to their learning. In particular, Group A is in general associated with a more desirable interaction and this could be the reason behind the positive correlation with learning. Although help-seeking frequency cannot determine learning effects on its own, the other aspects that were identified above seem to have the potential to play a role. All these will be taken into account in the next chapter during the attempt to model beneficial interaction.

4.6 Discussion

The qualitative and statistical analysis presented in this chapter, as well as their interpretation and implications were useful in several ways. First of all, in Section 4.3 the most important factors that could influence students' behaviours were explicitly discussed. In addition, a direct contribution to the field of AIEd was made. Results from educational research, which suggest that task orientation plays a role in students' behaviour in class, appear to play a role in the context of ILEs. Apart from contributing directly to this thesis, the results highlight a methodological issue for other researchers. Not only were the reported factors taken into account when designing studies, but the results justify the approach throughout this thesis of employing data-sets from realistic interactions with the system rather than artificial situations where the students are explicitly called to participate in a study.

The goal of Section 4.4 was to group students in an attempt to design appropriate interventions. The clustering process resulted in many clusters with too large a variance to allow the design of sensible individualised interventions or adaptation. However, it still helped in identifying overall patterns and choosing which replays to focus on and which students to call for interviews. Therefore, from a methodological point of view, this kind of clustering seems to be a viable approach to eliminate randomness when choosing illustrative cases to replay and students to interview.

The analysis presented in Section 4.5 contributed to the identification of individual differences in students' behaviours and the clarification of a definitely weakly-understood relationship between these behaviours and learning. Overall, it was identified that affective characteristics lie behind many aspects of students behaviour. This justifies, the theoretical underpinning of this thesis, and the fact that affect is pervasive in any educational situation. The analysis presented above also indicated that the amount of learning that occurs during early interactions with the system, has an influential role in students subsequent interaction and particularly their help-seeking behaviour. Students behaviours therefore has to be interpreted in the context of the rest of their interaction. The following aspects of students' interaction were of particular importance and provided guidelines for the redesign of the system and particularly the development of the machine learned models which are presented in the next chapter.

- It was established that there are several cognitive and affective factors influencing students' navigation and particularly the order with which they select pages, the time they remain on them and the fact that they abandon pages without completing their goals. For example, apart from the obvious fact that depending on an activity's difficulty students get discouraged or too confident, it was identified that they often do not appreciate the relevance or goal of an activity. Section 5.4.3 describes how the above influenced the redesign of the system.

- Students' response-giving process, and in particular the time they take to answer and whether they self-correct their mistakes, are good indicators of learning. The machine learning analysis, presented in Section 5.4.3, presents how these findings informed the development of a machine learned model which can predicts whether a student's interaction is beneficial in terms of learning.

- By investigating in detail students' help-seeking behaviour it was evident that they seem to ask more help than they need. Investigating the frequency of help-requests, a commonly used process measure in the field to drive feedback delivery, it was not trivial to differentiate between students who ask so much help that it becomes detrimental to their learning, and those who ask the right amount of help. The speed of hint-requests, whether students exhaust the available hints, whether they request the solutions, and the time they spend reflecting on hints seem to provide good indicators of the benefit of their interaction in terms of learning and were all used in the development of the machine learned models described in the next chapter.

Another outcome of the research process described in this chapter are some methodological issues. First of all, it was evident that averaging results across students and skills can often be misleading, depending of course on the system and its content. In WALLIS the skills vary significantly among them and therefore comparing groups of students which have interacted with different skills would yield deceptive results. In addition, in the case of the data used for the analysis of help frequency, the assumption of a normal distribution for conducting multiple regression did not hold. Unfortunately, it is not very clear if, such assumptions are always validated when reporting results from multiple regression, making a meta-analysis of controversial results among researchers more difficult. Finally, it was evident that results from traditional educational settings are not always transferable in ILEs. Only empirical studies like the one presented in this chapter can help to investigate students' interactions in more detail. Some of the aspects of students' interactions that emerged, would not necessarily have been observed in a controlled experiments.

Chapter 5

Re-designing the system

5.1 Introduction

Section 2.2 indicated that the methodology inspiring the current research is the Persistent Collaboration Methodology (PCM).

As outlined there, PCM is seen as a development cycle with many iterations which can start at any of the *observing*, *reflecting*, *designing* and *acting* phases. Results from this cycle spin off the development of theories as well as AI principles, techniques and tools. The PCM cycle of this research started with the observation of student actions in the previous chapter. In addition, determining patterns of interaction as well as establishing which of them seem detrimental for learning, signified the beginning of the reflection phase. The two phases lead to the usual outcomes of PCM according to Conlon and Pain (1996): (a) identifying the problem under investigation in a more concrete way, and (b) highlighting its complex nature. While this research is primarily interested in diagnosing students' emotions and their motivation, these are directly linked to the way the ILE under investigation works. Its details play an important role and can affect the research results. Inspired from the iterative nature of PCM and software engineering approaches which emphasise iterative and incremental design, it was felt that the aspects observed so far should be addressed before investigating issues related to affect. Behaviour is not independent of the environment in which it takes place, and similarly in an ILE, it is going to be detetermined as a reaction to the actions that the system takes. Therefore, investigating how student actions can be employed to diagnose affective states requires that the system is in a state that can engage students in what was described in Section 2.5 as an 'affective loop'.

This chapter presents the redesign of the system based on the findings of the previous chapter. The next section describes the principles underlying the redesign process. Section 5.3 presents two machine-learned models that were developed; the first predicts when students seek help unnecessarily, and the second whether their overall interaction is beneficial. The rest of the chapter describes the aspects of the system that were redesigned and the components introduced to monitor students' actions, and adapt its functionalities. For clarity technical details are avoided as much as possible and presented in Appendix E.

5.2 Redesign choices

Having in mind the need to investigate affective and motivational aspects of students' behaviour, several solutions were considered for the redesign, starting from simple ones to developing an animated agent. However, in the spirit of PCM, other research conducted on WALLIS (Abela, 2002) demonstrated that for this particular target group an animated agent was not so appealing. This is further supported by research in the animated agents field (e.g., Dehn and van Mulken, 2000) which shows conflicting results for their use. In addition, it seemed that the introduction of animated agents would open more issues (such as the 'persona effect' discussed in Lester et al., 1997) which are particularly intertwined with affect. Agents that are not well crafted tend to introduce more problems than they address (Moundridou and Virvou, 2002). Since affect was the focus of the rest of this research, such a radical change in the system would make students interact differently, introduce new variables in the analysis and would only complicate it more. Therefore, it was decided that, since the systems' effectiveness (at least for some students) was already established, it would be better to attempt to focus on adapting the current functionalities of the system with as few radical changes possible.

Two additional precepts were implicit in the redesign decisions. First of all, the overall goal of any redesign of an ILE should be to make it more effective by enabling it to provide additional support to the students who did not perform so well. Therefore, any approach should help them learn more. In addition, the redesign should not affect the students who were managing well with the system so far. These two precepts are similar to the two principles which Baker (2005, p.56) describes in terms of redesigning an ITS to adapt to students behaviour when they are gaming the system.

Given the importance of help-seeking identified in the previous chapters it seems that particular attention should be given to it. In Section 2.4.4 a model of good help-seeking behaviour as proposed from Aleven et al. (2004) was outlined. While extending the current model of interaction (described in Section 3.2.2.4) along similar lines of thought was considered, the interviews as well as early pilots of the first versions of the system raised arguments against this approach. Students were particularly concerned about feedback that interrupts their interaction with the system. As also described in Section 3.2.2.4 this was the main reason behind the feedback being delivered at the bottom frame of WALLIS. In addition, given the discussion about learner control in the Background chapter, directly refusing students help when they really need it can lead to undesired confusion, not to say frustration, with the system. The model in Aleven et al. (2004) is very intrusive and as they also recognise "not ready for live tutoring", particularly because it results in feedback being so often that interrupts students' work. Even in WALLIS where feedback is delivered without interrupting the student, the benefits of such an approach, which provides feedback for most of the student actions, are not so clear. Finally, given the inherent uncertainty of the information on which the system's decisions are often based, any approach should not be too restrictive in what it allows students to do, in order to avoid any negative effects in case a wrong decision is taken.

The aforementioned concerns extend to any approach that would tackle undesirable behaviour. The approach taken should not be too intrusive and interrupt students' actions. Baker (2005), facing a similar challenge when redesigning CMU tutors to deal with the 'gaming' behaviour, describes that the easy approach followed in this kind of situation is to tackle the undesirable behaviour by choosing a *preventative* approach. This means changing some characteristic of the system in order to prevent the observed behaviour. This was also the initial approach in WALLIS. Further hinting was prevented by not allowing students to ask for solutions if they had not tried first to answer the question. This however, as described in the previous chapter, led them to answer randomly and request hints quickly just to enable them to request the solution. Similarly, Murray and VanLehn (2005), describe a preventative approach that led students to develop new strategies for gaming the system by rapidly repeating the same error several times so as to enable the system's proactive help. Even though this behaviour could also be dissuaded, as Baker (2005) also observes, this leads to an 'arms race' where students are developing harmful (in terms of their learning) interactions and the designers are trying to stop them. Apart from the fact that it is not optimal, these changes of the system are antithetical to the principles discussed above and particularly that of not making changes that would interfere with the way students were working so far and learnt from the system. Based on all the above, wherever possible less preventative solutions should be preferred to any preventative ones.

In addition, Section 4.5.3.2 established that help-seeking frequency explained only some of the variance in learning when prior ability was partialled out and also only under certain circumstances. Throughout the chapter, other factors such as the speed of answering the question, the time spent between consecutive hints, the time the students reflected on the solution or the hints they received, seemed related to learning.

Given the above, it seemed unclear at this stage how an approach that focuses on help-seeking behaviour could be designed in a way that guarantees its effectiveness. However, based on the results of the previous chapter it seemed possible that a measure of 'desirable' interaction could be defined that, without intervening a lot, could empower the system with an indication of the students' benefit from the interaction so far. Based on this measure, students could be provided with additional practise for the skills they have not adequately practiced. As a result, it was decided to develop a model that would have the ability to predict if students' behaviour is beneficial to their learning. The model could guide further adaptation of the content, the feedback provided as well as interventions of the system. Throughout Chapter 4, it was also made clear that the amount of learning that occurs during students' interaction with previous relevant items, plays an important role not only in the correctness of the answers provided, but also in their help seeking behaviour and some affective states. This suggests that being able to predict how the interaction with prerequisites affects the need for help for a certain step would also be very useful. The value of such a prediction is further supported by tutors' comments when interacting with students over the computer-mediated environment described in Section 2.5. One piece of evidence elicited from the tutors, was the expected effort of the student in relation to their knowledge and probability of knowing the skill or having learnt it during the session (Porayska-Pomsta et al., 2008). Based on the above, a first step in developing a model of *beneficial interaction* was to develop a model that predicts if a student could have acquired from the system the necessary knowledge to answer correctly without any *need for help* from the system. These two models are presented in the next section.

The rest of the redesign choices are based on the implications identified in Chapter 4 in relation to students' way of providing answers, navigating and quitting pages. These are presented in Section 5.4.

5.3 Modelling students' interaction

5.3.1 Predicting the need for help

The goal of this section is to derive a measurement that provides a prediction of the ability of students to answer a question correctly without any need for help. The benefits of this prediction are twofold. First, in general, it is a useful prediction on its own for the design of appropriate interventions (e.g., provide unsolicited help). Second, in this research, it was used as a feature in subsequent parts of statistical analyses; not only in the model of beneficial interaction but also in the affective diagnosis rules (in Chapter 6). Apart from this short-term goal, the long-term goals are (a) to use this model, and the methodology for building it, in a general way in conjunction with other evidence-based frameworks (for example the one by Morales et al. (2006) presented on page 27) but also (b) potentially, to transfer the methodology (and possibly the results) to other courses using WALLIS or optimistically to other systems.

The problem is not unique to the current research. Many ITS, which try to adapt their feedback accordingly, need such a prediction. However, it is not easy to measure the exact effects of a student's interaction with the system. The individual differences between students, together with other affective characteristics seem to determine whether they request help or not. As the problem is quite complex, different researchers address it in different ways depending on the special characteristics of the system and the overall context. For example, in the CMU tutors (e.g., Anderson et al., 1995) the problem is approached as an attempt to estimate the probability of knowledge that the skill has been mastered; a technique they call *knowledge tracing* (see Corbett and Anderson (1992)). Similarly, (Martin and Vanlehn, 1995; VanLehn and Martin, 1998; Conati et al., 1997) describe systems where Bayesian networks are used to predict students' knowledge during the interaction with the system. To reduce the complexity (and to avoid duplication of work) the issue was approached slightly differently in this research. The approach taken is to predict whether a student needs help on an item based on their interaction with previous parts of the system. Rather than employing arbitrary rules or models based on intuition, an attempt is made to derive a model from data.

As discussed in section 4.2.2 the GIC data-sets were collected from studies where students come with no background knowledge, so the material and the way they are taught are completely new to them. Therefore, it does not seem too bold to assume that students who do not ask for help and answer correctly with the first attempt at a question have learnt either from carefully reading the material in the system or from the interaction with the related exercise. Therefore, all other characteristics of a student being equal, similar interactions should have provided the student with the necessary skills to answer without the need for help. The opposite is not necessarily true. It has been established already that students ask for hints for different and complicated reasons.

1. *trp* : time spent on related page
2. *tsa* : time spent on attempt
3. *prev:* previous knowledge as indicated by their performance in the prerequisite course and assignments
4. *rel:* a rule-based measurement of the degree of 'completeness' of the goals of interactions on related pages
5. *diff:* difficulty of the item
6. *answertype:* the type of the answer required (mcq,blank,matrix,checkbox)

Table 5.1: Possible variables for the model of predicting the need for help

Initial investigations with the GIC03 dataset as a learning set and the GIC04 as a testset, supported the claim that such a model could be used to automatically predict with reasonable accuracy (more than 65%) the need for help. It was decided to focus the prediction only on help requests before the first attempt to answer a question. Any further interaction, after the students' initial attempt, is quite complex and therefore was ignored for the model construction. As the previous Chapter demonstrated, further help requests often depend on students' understanding of the feedback, whether they read it or not and several other factors, which complicate the prediction task.

Given the above assumptions on what exactly to model and in order to learn a more accurate model from the data both the `GIC03` and `GIC04` datasets were used as a training set. Vectors were constructed that contain the variables shown in Table 5.1 and a binary class that takes true and false values. In order to have a simple and generalisable model as few features as possible were used. Also an attempt was made to employ information that is considered accessible in other systems. The set of features that were initially considered for the model are shown in Table 5.1. The class learned represents whether the student seems to require help in order to answer. Its value therefore, is TRUE when students provided completely wrong answers (not from usual misconceptions), or answered very quickly[1] demonstrating, in a sense, that they did not read the question carefully and they may be answering randomly just to get feedback. The value of the class is FALSE when a student's answer is correct or partially correct (according to a list of common misconceptions - not applicable in MCQs). Students who asked for help without an attempt are not included. The main rationale behind the latter is that, as was discussed in Chapter 4, there are many explanations behind the request for help and using these data for the machine learning does not necessarily provide instances that demonstrate whether a student really needed help or not. All the above restrictions resulted in a set of 1230 attempts (429 of which were unsuccessful).

The next step was to choose the exact modelling approach. Preliminary investigations with cross-fold validation with the combined GIC03 dataset suggested that from all the approaches attempted (decision trees, Bayesian network, classification via regression) the Bayesian network and the decision tree were the most accurate ones and very close to each other. The Bayesian network was therefore preferred mainly because of the uncertain nature of the prediction.

To learn the network, the ICS algorithm of Weka was employed as described in Section 3.4.3. ICS is the only algorithm in WEKA that claims to attempt to learn causality. The conditional independence tests of ICS left out the variable `answertype` from the model as irrelevant. Fast Correlation Based Filtering (described in Section 3.4.4) also confirms the relevance of all variables apart from `answertype`. The final model learned

[1]The usual discretisation described in Section 3.4.6.2 was followed

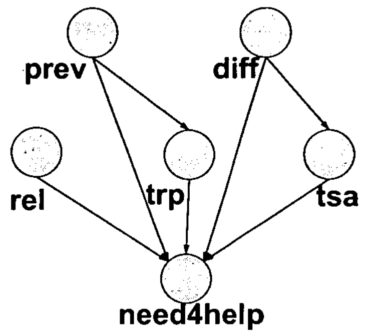


Figure 5.1: Bayesian network for predicting need for help

appears in Figure 5.1.

To evaluate the result the `GIC05` dataset was used as a testset. The accuracy report is presented in Table 5.2. The recall measurement (also discussed in Section 3.4.5) indicates the ability of the model to select instances of a certain class from the data set. Apart from the encouranging high accuracies, of particular importance is the high recall of instances classified as FALSE (i.e. where the model would predict that the student does not need help). The high values indicate less false negatives (i.e. less cases where students need help but the model predicts that they do not). The model's accuracy was considered adequate for the purposes of the research here. This has to be interpreted in the context of the likely educational consequences of incorrect predictions given the way it is integrated with the rest of the system. This is discussed in Sections 5.4 and 5.5.

Further investigation with the data showed that considering different models for every item (as shown in Table 5.3) improves the results substantially. The main reason behind this is the fact that some of the variables have different effects for the different items. For example, in some items the related exercise does not play such an important role. Therefore, one model could not accommodate all the items. To construct the different Bayesian networks the data were split according to item. The variables `difficulty` and `answertype` of Table 5.1 were removed since for every item these values were static. This simplified the models considerably. Further simplifications occurred from the conditional independence tests without affecting the average accuracy.

		BayesNet		J4.8	
		Cross	Test Set	Cross	Test Set
	accuracy	67.64	66.52	65.84	64.05
	Kappa	0.317	0.318	0.30	0.23
recall	True	0.60	0.56	0.59	0.5
	False	0.74	0.76	0.71	0.72

Table 5.2: Classification accuracy, Kappa statistic and recall values for two different classification techniques to predict answering correctly without need for help.

item id	description	related item(s)
GD-A	interacting with applet	CD
CC-G	classifying conics from graph	CD, GD-A
CC-E	classifying conics from equation	CC, GC-G
SF-P-AM	find associated matrix	SF,SF-EX
SF-P-EG1	find eigenvalues	SF,SF-EX, SEQ(ga)
SF-P-SF	find standard form	SF,SF-EX, CC-E
SF-P-SF-C	classify the conic from SF-P-SF	CC,GD,CC-E
S-P-SF-EG2	find the eigenvectors	SF,SF-P-EX, SF-P-EG1
S-F-SF-ROT	angle of rotation	SF-P-EX, SF-P-EG1, SF-P-EG2

Table 5.3: Items and related pages and prerequisites (see Appendix B for the material)

Therefore, the separate models were preferred for the actual implementation.

Moreover, driven by the high accuracy of the individual models an attempt was made to lift the restriction of modelling only items where the student has no prior knowledge. This way the need for help in the two items on which the students have prior knowledge (`SF-P-EG1` and `SF-P-EG2`) could also be modelled. In this case the variable `rel` described in Table 5.1 was changed to reflect the result of the assessment in the exam of the related course; Solving Equations (SEQ). As discussed before, a particular question in the exam was probing students' understanding of eigenvalues and eigenvectors. If the student was not successful in the exam then the variable reflected

		BayesNet		J4.8	
		Cross	Test Set	Cross	Test Set
	accuracy	69.12	67.61	67.84	63.05
	Kappa	0.36	0.35	0.34	0.27
recall	True	0.74	0.71	0.72	0.66
	False	0.62	0.62	0.60	0.58

Table 5.4: Average Classification Accuracy and Kappa statistic for Bayesian networks and decision trees to predict answering correctly without need for help

the interaction with the system as in all other cases. If the student was successful in the exam and only if they interacted with the relevant pages then variable *rel* reflected whether their interaction was beneficial[2]. As expected, it turned out that these two models were less accurate than others (around 64% and 65%). This is quite reasonable as there are many other factors that play a role in this process and are not included in the model. For example, the results in SEQ could have been an artifact, a circumstantial result, or simply the student could have used other means to revise this particular part. In addition, long term retention of the knowledge is not necessarily reflected with the general ability of the student. Therefore, it is hard to have accurate data for all students. These problems are reflected in the accuracy of the models. On the other hand, their accuracy is satisfactory at this stage since they will be used to inform the Bayes network for predicting whether their interaction is beneficial. In addition, as discussed in Section 5.2 keeping in mind the uncertain nature of such predictions any feature of the system that takes them into account should be designed accordingly. The errors in the above predictions justify the choice of not designing a preventative approach that would not allow students to ask for help and rather take their behaviour into account for a prediction of how beneficial their interaction is.

The average of the results for the different models is presented in Table 5.4. Note that the averages presented include the two items with the low prediction. Without

[2]As described in section 3.2.2.4 the system did not present the relevant pages by default, but only if requested by the students. In addition, students who answered correctly in the exam tended not to ask for help in this particular part.

them the average is much higher (68.37%). Detailed results and the actual structure of the networks appear in Appendix D. The Appendix also shows the accuracy of logistic regression, attempted retrospectively, after the implementation of Bayesian networks in the system was complete. Although logistic regression seems slightly more accurate in certain cases (on average it has accuracy 68.916% against the testset and for the single model 69.887%), for the purposes of the research here the slight improvement in the model was not considered significant for immediate implementation but future work will investigate this issue further.

The aforementioned results suggest that the above model could be used adequately for predicting whether a student needs help or not. The separation of the model by item seems to be against the long-term goal of coming up with one model that could be used in other lessons and/or systems. However, the methodology for building the models can be used in future work. In addition, the approach followed here does not require human intervention and could be automated allowing the system to learn and improve while it is used. Since this was not the main focus of the research conducted here but just an intermediate step the issue was not investigated further. In addition, although there are several ways the result of the overall model could be improved, it was more important at this stage to have good accuracy not only because of the need to adapt aspects of the system but also for the modelling and prediction tasks that will follow in Chapter 6. Therefore the multiple models were preferred at this stage. Chapter 7 provides thoughts on how to improve the accuracy of the overall model.

Finally, it was established that the initial assumption of modelling only items on which students do not have background knowledge was not necessary. It seems that despite the lower accuracy it is possible, by using information from previous courses, to derive an adequate prediction. In the case of WALLIS the availability of previous results in the prerequisite course were enough. In other systems this information can come either from interaction with the system in advance or by entry tests. It is worth saying here that as data was that continuously selected the model, it can be improved further and, as discussed above, this process could be automated, in the future.

5.3.2 Beneficial interaction

This section describes the development of a machine-learned model which predicts how beneficial a student's interaction is. The assumption behind the attempt to build the model by learning from data is that the behaviour of the students is associated with their learning effects. Another implicit assumption is that differences in learning style, in affective characteristics and other preferences are reflected in students' behaviour and unravelling which behaviour leads or not to learning gains can provide a useful mechanism for improving the feedback provided by the system. One way to approach the issue was by utilising the `GIC04` dataset which includes learning outcomes. If a model, which predicts these outcomes satisfactorily, could be learnt, then it could be used to guide interventions or feedback provision.

Before choosing the exact modelling approach, a set of possible variables was derived. These could be used as features in the learning task. Based on the investigation with the data the variables that appear in Table 5.5 were considered.

For similar reasons as the ones described in the previous section, Bayesian networks were preferred. First of all, informal comparisons with decision trees established that they had similar accuracy. Comparisons with regression showed that they were superior. In addition, the nature of the prediction is again highly probabilistic making it a perfect candidate for Bayesian networks together with the fact that, in future implementations, this information could be part of a larger evidence-based probabilistic frameworks.

To learn the Bayesian network we employ again the ICS algorithm of Weka and to facilitate the algorithm's search FCBF is also employed in advance to remove irrelevant and redundant features. Although (as expected) the results were very similar (i.e. the same more or less accuracies and structure was learned with ICS or Hill Climbing over the complete set of features) simpler models are always preferred (Occam's razor). In fact, the simplified model achieved better accuracy on a 10-fold cross validation check and slightly better accuracy on the test set. By removing redundant features the remaining ones were easier to comprehend. This allows a more sensible ordering of the variables, which as discussed in the introduction, can effect the search for the structure of the Bayesian network. Finally, the process with FCBF was significantly faster. In

1	Help frequency.
2	Error frequency.
3	Tendency to ask for help rather than risk an error (as defined in Wood and Wood, 1999) $\frac{help}{errors+help}$
4	No need for help (according to Section 5.3.1) but help requested (true/false)
5	Answertype - the type of the answer required (mcq, blank,matrix,checkbox)
6	Previous attempts in items related to the current skill If this is the student's first opportunity to practice this skill: -1 If no previous attempt was successful: 0 Otherwise the probability predicted from the model in previous attempts.
7	Whether the prediction was beneficial in related items (or high mark at the exams of the prerequisite course if no related items are on the system) (true/false).
8	Time in standard deviations off the mean time taken by all students on the same item.
9	Speed between hints – The Mahalanobis distance of the vector of times between hints from the vector of mean times taken by all students on the same hints and item.
10	Accessing example while answering (true/false)
11	Self-correction (true/false)
12	Requested solution without attempt to answer (true/false)
13	Reflection on hints (defined as the time until next action from hint delivery) (calculated similarly to 9 using again the Mahalanobis distance).
14	The number of theoretical material lookups that the student followed when such lookup were suggested by the system (-1 if no lookups were suggested)

Table 5.5: Features considered for learning the model of beneficial interaction

1.	Tendency for help
2.	Need for help
3.	Self-correction
4.	Example access
5.	Average reflection time
6.	Speed for hints
7.	Error frequency

Table 5.6: FCBF selected variables from Table 5.5

		BayesNet		J4.8	
		Cross	Test Set	Cross	Test
	accuracy	70.112	68.234	66.517	65.843
	Kappa	0.401	0.364	0.318	0.313
recall	True	0.726	0.726	0.714	0.686
	False	0.672	0.623	0.605	0.626

Table 5.7: Classification accuracy and Kappa statistic for two different classification techniques to predict beneficial interaction

this case, since all the analysis was performed asynchronously, and the models were implemented in advance this was not really relevant. However, in case the techniques reported here are automated to enable the system to learn while more students work with it, online speed will become a more critical factor. The list of reduced variables is shown in Table 5.3.2 and the final model in Figure 5.2.

In contrast to the previous example deriving different models for every skill did not improve the accuracy of the models substantially. This is probably because, in this case, the variables do not play different roles in each different model and therefore the quantity of the data increases the accuracy for the learning task of the whole model. The accuracy of the model, as well as comparisons with decision trees are presented in Table 5.7.

Although, as in any model, further investigation and research could improve its accuracy, the model was considered adequate for the purposes of the research here. More

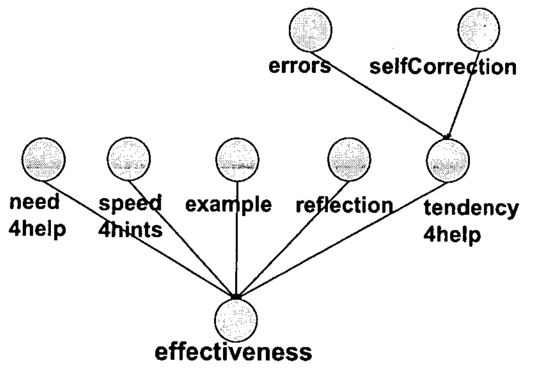


Figure 5.2: Bayesian Network for predicting beneficial interaction

detailed comparisons with J4.8 appear in Appendix D. As before, logistic regression was considered after the implementation of the Bayesian models. Since it has slighter better accuracy (70.337%), it seems that in future work logistic regression could be used instead of the Bayesian network for the implementation. Logistic regression not only can be implemented more easily but it can also provide the probabilistic framework needed to deal with the uncertain nature of the predictions here.

The next section describes how the two models developed here were employed at this stage of the redesign. In Chapter 6 the outcomes of the models in each step of the user interaction are used as a feature for the prediction of affective and motivational states. Other suggestions for improvements and useful places were these models could be used are discussed in the last Chapter.

5.4 Architecture of the re-designed system

The redesign of the system was centred around the feedback mechanism of the original system (Section 3.2). As already discussed in Section 5.2 the approach taken should not be too intrusive (e.g., interrupting students' work in order to provide feedback) nor preventative (e.g., preventing them from asking help). The accuracy and the probabilistic nature of the prediction further justify these redesign choices. In particular, if one takes into account that in around 30% of the cases the model could be wrong it is obvious, but paramount, to use these predictions in a way that has the fewest negative educational consequences. The main implication of this principle in the re-design of the system was the fact that the prediction for necessity of help-requests was not employed directly but rather through the model of beneficial interaction. The following subsections demonstrate how this was taken into account in the re-design of the feedback mechanism.

In order to maintain the separation between components of the system one component was added called the *modeller* that employs the Bayesian networks above and can predict whether the students' interaction after an exercise is beneficial, and whether their help-request on a step was superfluous. The outcomes of the modeller are taken into account by the feedback mechanism in order to adapt the system's actions. In addition, the component dealing with navigation help which was previously part of the overall feedback mechanism was separated and its functionalities were changed by taking into account current standards and techniques (see XML in Appendix B) in order to provide more elaborate suggestions and to be able to link with the output of the modeller. To facilitate that, changes were made to the component that records students' interactions. The enhanced overall architecture is depicted in Figure 5.3 while the necessary changes are explained in the following subsections.

5.4.1 Interaction capture agent - *logger*

As described in Section 3.3 this agent was already recording every aspect of students' actions for the purposes of the research in this thesis. The agent was adapted to perform summarisation of the information required for the machine-learned models of the

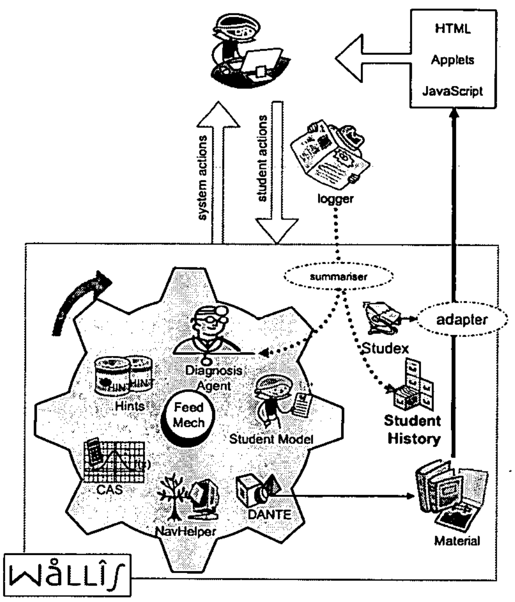


Figure 5.3: The redesigned architecture (figure adapted from the original architecture of WALLIS, Figure 3.2)

previous section in order to optimise it so as to transfer less data over the network when full replays of the interaction were not needed. Accordingly, the following directly observable aspects of the interaction were recorded for each item with which the student interacts:

- time spent on the item
- hint requests and the speed they are requested with
- type of errors on steps and self-corrections before submitting answer
- whether the relevant example of the item was accessed
- hint reflection

In addition, when leaving a page a prediction request is made to the *modeller* (see next subsection) so as it can be recorded for future calculations. Finally, the reasons for leaving pages (see prompt in 5.4.3.2) are recorded.

5.4.2 Diagnosis agent - *modeller*

The diagnosis agent is server side and contains the models described in Section 5.3 uses the evidence logged from the interaction capture agent constantly calculates

- probability of need for help
- the probability of the interaction being beneficial

These were built using JavaBayes[3] In addition, when requested from the *logger* predictions are forced and saved for future reference. Finally, the page is annotated for completeness only when the model predicts that the interaction is beneficial. Finally, the agent keeps a tree structure of the material derived from an XML file (see Appendix A) in order to know which pages are related with which,

- tracks students' learning path in an attempt to allow freedom of choices and the feeling of more locus of control to the student. As discussed in Section 4.5 the results from the exploratory analysis suggest that students often just visit pages to see what they entail in order to come back later. This should be taken into account. Most systems not only treat the visit of a theoretical of example item/page as evidence that the student read the page, but also that this implies something for their knowledge (eg. . Even with such tools such as the Poor Man's Eyetracer (Ullrich et al., 2003). In WALLIS therefore part of the metadata of the page we introduce a time in relation to the typical student reading this page before taking it into account in the student model .

5.4.3 Feedback Mechanism

The feedback mechanism (described in Section 3.2.2.4) was enhanced to take into account the output of the models above, as well as the implications drawn throughout Chapter 4. Apart from changes to the local exercise feedback, the mechanism was improved with the ability to provide suggestions about which pages to study (often referred to as *global feedback* (Melis and Ullrich, 2003; Eugenio et al., 2005)) but also

[3] http://www.cs.cmu.edu/ javabayes/

the reasons to interact with them. In addition, the prompt described in Section 4.5.1.3 was also enhanced providing the system with the ability to gather more evidence on the reasons the students are abandoning a page.

5.4.3.1 Navigational Feedback

In Section 4.5.1.4 it was explained that (particularly in the context of higher education students) it is important not to restrict students' sense of locus of control. Given the absence of evidence that following a particular order of the material was more beneficial than not, it was hypothesised that not intervening at all and allowing students to interact based on their preferred learning style should be more beneficial. While this requires further testing the current redesign addresses the issue by following the usual tenet behind ITS which was discussed in Section 2.4 that of allowing the decision about when to seek help to the student.

In the first place, all navigational hints or links to learning materials are changed to include the relevance of the material towards the students' goal. For example, instead of simple hints like "`Have a look at this page`" the link is more explicit and some rationale is given. For example, "`It looks like you haven't seen the first Classifying Conics exercise, it will help you practise before answerring here.`". This is based on a short description of the goals of each item that are predetermined in advance in a tree like representation based on the representation of the goals in Figure 3.3 on page 39. The XML structure that supports this, which includes the goals of each of the items, is presented in Appendix A.

In the second place, an additional level of help is provided to the students when they ask for global help:

- During an interaction with an exercise, if the students asks for global feedback, the diagnostic agent performs a prediction of how beneficial the interaction with the item is so far. Then:
  - if this is low or medium then students are reminded of the goals of the exercise and are encouraged to complete the whole exercise.
  - if this is high then a comment similar to *'It seems you are managing quite well/very well'* is provided followed by an encouragement to complete all

the remaining items before moving on to the next activity. A link for the next activity is also provided.

- If they have just completed the interaction then
  - if the *modeller* predicts that their interaction with the item was beneficial, suggestion for the following item in the list of goals is provided together with a short description of its goals
  - if not, the system suggests that the student tries the exercise again.

5.4.3.2 Page 'abandonment' feedback

As described in Section 4.5.1.3 the prompt that was introduced in the early versions of WALLIS which was warning the students that they did not complete the page, apart from having the effect of reducing the page abandonment level was also received positively by the students. They were either on the course of leaving the page already (so it did not necessarily obstruct them) or the question was only helping them in either staying there (if their attempt to leave was accidental) or was allowing them to request more help by asking for an example or theory. Given the above the prompt was adapted to ask the students the reason for leaving the page (see Figure 5.4). Therefore although technically the prompt is part of the feedback mechanism it also communicates with the Interaction Capture Agent component in order to record the reason (see Figure 5.4).

The prompt appears only if the goals of the item were not achieved in the past at least once. In addition, given students' comments during interviews and observations of their interactions, it does not appear if they leave a page immediately after visiting it. The exact text of the prompt is adapted according to the prediction of beneficial interaction. More specifically, if the modeller predicts a value below medium the prompt says `"Instead of giving up perhaps you may need some help. Why don't you ask for some hints or read the example again/first"`. If, however, the prediction is high it says `"Based on your interaction so far it looks like you could manage on your own"`

The introduction of the prompt could enable more informed input for the diagnostic agent for future work. That is, if the example or theory is accessed then it could be con-

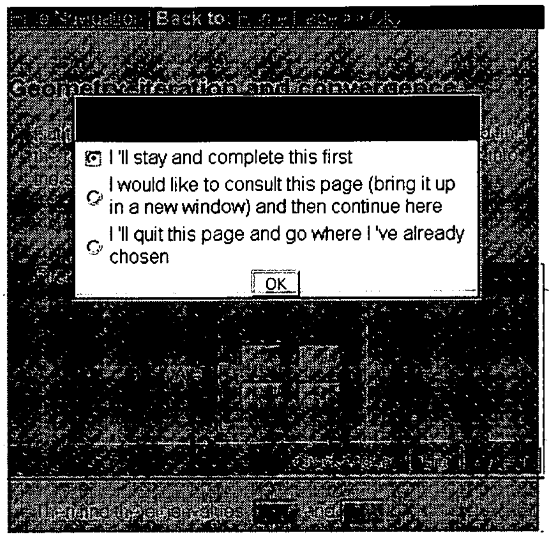


Figure 5.4: The prompt that appears when students are attempting to leave an item. The content is blocked and the students have to make a choice before proceeding.

sidered as part of the help-seeking behaviour and treated by the appropriate component (i.e. the system could check if this help is necessary based on its prediction). Students' behaviours with the prompt could be used as evidence for other aspects. For example, affective characteristics or the model of beneficial interaction. Also, in Chapter 6 this information is used as a feature for the prediction of affective characteristics.

5.4.3.3 Providing answers and requesting help

These two aspects of the interaction were not changed much. As explained in Section 5.2 a non-preventative approach was chosen. Rather than disallowing any help or solution requests these are taken into account at the model of beneficial interaction.

An attempt was made to address the observations in Chapter 4 about random and copied answers. Although predicting random answers is left out of the scope of this thesis, as it would constitute a research topic on its own, they are indirectly tackled by the fact that the speed of answer was taken into account in the model of beneficial interaction. At this stage, only definitely copied answers generate explicitly negative feedback suggesting to the student to reconsider the answer.

The only other change that occured to the feedback provided during an exercise is after a hint request and when a misconception in the last provided answer is detected. If the interaction of the related material was not beneficial then the first hint provided suggests to the student to visit the relevant page. The goal of the suggested page is also explained (c.f. Section 5.4.3.1).

5.5 Discussion

This chapter described the redesign of the system based on some of the results observed in Chapter 4. PCM provided a helpful methodology in the sense that through the observation phase and the beginning of the redesign phase, it was made clear that any solution adopted to tackle affect would be partial as it seemed that, apart from cognitive aspects, the identified behaviours were related to affect. Without redesigning the system to address at least some of the aspects observed, conducting studies to investigate issues that relate both to student actions and to their affect would be futile. By designing a more appropriate interaction with the system, and gradually engaging the students in a more 'affective loop', the system is at a stage were affective diagnosis could be incorporated to adapt the feedback and interventions further. According to the tenet of this thesis, that of designing based on clear empirical evidence and not on assumptions, this could not be done without further studies.

The re-designed system was integrated in the classroom for the `GIC` module in 2005 and thereafter. As mentioned in Section 4.2.3, the lecturer teaching the course was not the same as in the previous years and did not feel comfortable using the re-designed system as the sole means of delivering the material. Therefore, `GIC05` was considered only as a good opportunity to collect more data and conduct the evaluation of the machine learned models. In addition, the technical efficiency of the *modeller* and the changed feedback mechanism were tested by simulating their outcomes in a log file. Although a more formal comparison between the different versions of the system would be useful and necessary, ethical and practical considerations hindered such an analysis. However, the system and particularly the feedback mechanism worked as expected. Since the re-design choices were arranged not to prevent students from taking actions, it can be assumed that by its design, the feedback they would receive is of higher quality, since it is more adaptive to their interaction. The evaluation of the machine learned models suggests that the prediction about beneficial interaction can be wrong around 30% of the times. However, even in the case where the prediction is wrong, students just receive encouraging prompts without specific feedback (for example, a suggestion to continue with another page rather than suggesting to work more on the page). The negative educational consequences of the above strategy are probably no worse than allowing students to abandon a page without any feedback. The next chapter describes a study which was conducted with the re-designed system integrated in classroom in 2006. No negative aspects were observed or reported by the students but their interviews indicate that there could be affective and motivational implications of feedback strategies such as the aforementioned. In the long-term, continuous feedback based on wrong predictions could send implicit messages to students, for example, about the way they are expected to interact with the system, perhaps promoting a performance-oriented learning approach. Further work should investigate this issue more formally.

Chapter 6

Predictive modelling of affective states

6.1 Introduction

The research described in the previous chapters and the redesign of the system, apart from serving their respective goals, were preparatory to the investigations of the possibility of detecting affect based particularly on students' interactions with ILEs. As was discussed in the Introduction, the use of students' direct behaviour with the system to predict emotional and motivational characteristics is important. While using facial, or other bodily measures could be potentially more reliable, observable actions constitute the way students interact with the system in any case. Therefore, the technology required to collect them is less obtrusive and less expensive. In addition, recording actions as they take place in the system, does not require further student involvement (e.g., a camera or a glove) and therefore, the students' feeling of being monitored is reduced. Moreover, as discussed in Section 2.5, the issue of employing student direct behaviour with the system as predictors for students' affective and motivational characteristics has not been investigated in detail. The little research that exists looks into the issue either based on a theoretical approach (e.g., Jaques and Viccari, 2007) or employing a more traditional knowledge elicitation technique based on a qualitative analysis of expert interviews (e.g., de Vicente, 2003).

The primary goal of the research presented in this chapter was to develop a methodology for collecting and analysing data in such a way that machine learning can be employed to enhance the results of qualitative analysis. A secondary goal, but equally important, is to verify that the affective characteristics, which related research suggests that occur during learning (see Section 2.3), are relevant to the context and situation of WALLIS and to investigate which of them can be detected in the mode of interaction that is assumed behind ILEs similar to WALLIS.

Towards these goals, Section 6.2 describes a study conducted with the aim of collecting data for modelling students' affective states and the results of the machine learning analysis performed based on their verbalisations about affect when watching replays of their interaction with the system. Although the results are satisfactory, Section 6.3 discusses the limitations of the approach and in particular the fact that modelling based only on data from self-reports can result in a biased model. In addition, regardless of how valid the diagnosis produced in this manner is, it is not clear to what extent it is useful in terms of facilitating the adaptation of the tutoring. It seems therefore equally important to take into account tutors' inferences about these factors since it is them who are called to adapt their teaching accordingly (Porayska-Pomsta et al., 2008). Section 6.4 discusses this in detail and outlines two different approaches followed in order to take into account tutors' perspective.

The first, presented in Section 6.5 attempts to elicit data by asking tutors to infer students' affective states during replays of students' interaction. A pilot study met with difficulties which stem particularly from the fact that the tutors were asked to diagnose a situation in which they were not involved directly. This suggested that a different kind of study where tutors would be the ones providing feedback and suggestions to students would be more appropriate. Although such a study was not possible in the context of this thesis, the second approach, presented in Section 6.6, was to analyse a dataset which consists of tutors' affective diagnoses collected from the LeAM empirical studies (mentioned on page 28) with tutors and students interacting with each other in an environment approximating an ILE. Appendix A shows the chronology of these studies.

While the machine learning analysis performed provided useful results, some problems are discussed in Section 6.7 that demonstrate that collecting tutors' inferences is not necessarily trivial. In the light of the difficulties associated with both students' and tutors' externalisations the data collected here serve mainly as means of investigating the machine learning methodology behind their analysis. The results themselves serve mainly as indicators for future work.

The issues associated with the machine learning analysis of these data are presented in the respective discussion section of each study. To avoid repetition, more general issues that pertain to both studies are discussed in the final chapter. In all the machine learning analyses performed in this chapter, apart from straightforward statistical techniques to investigate the data, decision trees were employed. The general advantages were described in Section 3.4.1. It is worth repeating that, apart from the fact that they can be easily converted to rules and therefore implemented straightforwardly (for example in the Jess component of WALLIS), they provide results of which the form is human inspectable. This makes it easier to interpret the results. In addition, at this stage where there is no substantial evidence to support the generated rules, these rules can be further validated employing qualitative methods (e.g., revisiting the data) and can be used to generate hypotheses for further research.

6.2 Predictive modelling from student perspective

This section presents the study, its methodology and the analysis conducted in order to approach the goals described above. The main goal of the study was to collect data that would facilitate the prediction of students' affective states from their interaction with the system. Section 6.2.1 describes the methodology of the study, its context, how participants were chosen, the process followed and particularly the choices in relation to data collection. Section 6.2.2 provides a brief descriptive analysis of the results that facilitated the machine learning analysis. The methodology of the latter is presented at Section 6.2.3 and the results at Section 6.2.4. Finally, Section 6.3 provides a discussion of the issues closely related to this study. A discussion of more general issues is presented in the next chapter at Section 7.3 which addresses issues both from this study and the next ones described in the rest of the chapter.

6.2.1 The methodology of the study

6.2.1.1 Context

As discussed elsewhere (Sections 2.1, 2.4.3) students' behaviours and affect can be greatly influenced by the general context of the study. In particular, the goals and beliefs of the situation under which students were asked to use the system play a significant role (see Section 4.3). In addition, it was established in previous studies (for example the exploratory one presented in Chapter 4) that cleaner data are collected when students interact over a substantial amount of time with the system. As a result, the GIC course (see Section 4.2.2) was employed once again. Especially in this study that investigated affective characteristics which are more easily influenced, particular attention was given in the way the system was introduced. First of all, the students knew from the beginning of the course that they were expected to use an ILE anyway. In addition, most of the students had used the particular system before.

According to the principle of interfering as little as possible with the way students interact, and the explicit effort to maintain ecological validity (as described on page 10) students were allowed to work in their own time and location rather than setting up a specific environment and asking them to participate to the study. However, a small pilot test with seven students was conducted prior to this study were students interacted with the system in a lab while their interaction was recorded[1]. This allowed further testing and tuning of the recording agent and informed the design of the questionnaire, which is described later. In addition, it helped determining what resources, particularly equipment and time, are needed to replay interactions to a student and refining the exact issues that would be addressed during replays of their interactions (also described below). In particular, the pilot helped in establishing how to communicate with students the needs of this research and what could be expected from them in relation to their understanding of the affective factors. The findings are mentioned in their respective sections below.

[1]In fact the pilot was a fallback of a pilot that would look into deriving the same information from involving students and tutors in a 'wizard-of-oz' type of experiment . While students were lined up it was established that finding appropriate tutors was quite difficult for this context, required a lot of resources and a lot of their time and therefore never took place.

6.2.1.2 Selection of participants

The participants were selected from a group of 209 undergraduates who attended the 2006 session of the "Geometry and Iteration" course[2]. Budget and time constraints for this part of the research permitted the involvement of about 20 students only. Therefore, a representative sample was needed. This sample had to include all subgroups of our interest (i.e. students with different previous abilities and self-awareness).

Groups were formed based on students' entry qualifications and their mark in a prerequisite module (and particularly a question that relates to skills prerequisite to the material covered under the study). This method of selecting students is similar to a disproportionate stratified random sampling (Lohr, 1999) and assures that any statistical results will be more precise than when selecting a sample randomly. Initially 23 students volunteered across the different groups. As we were interested in having at least some students that are (or think that they are) familiar with the concept of conics a further selection took place using a pre-test.

From the selected students 18 (3 with very low, 5 low, 5 medium, 3 high and 2 very high previous performance) completed all necessary tasks and faced no technical problems. These students were also administered a post test which examined the knowledge acquired by interacting with the system.

6.2.1.3 Pre and Post Tests

The pre-test consisted of 10 questions (see Appendix B) targeting knowledge and familiarity with conics sections . As it was known (from the research with the previous data-sets and the lecturers' experience) that most students have no (or in the best case little) prior knowledge, the point of the pre-test was not so much to establish their knowledge in conic sections but mostly to establish which of the students have had some contact with the subject before and thought that they knew something. The first five questions at the beginning of the pre-test were particularly targeting this.

Only 3 students (1 in the medium group and the 2 of the v.high group) claimed to be familiar with the existence of the concept of conics but like the rest of the students

[2]The course was renamed from "Geometry, Iteration and Convergence"

failed to answer any question of the pre-test. The post test (6 questions) targeted particular aspects of the material and knowledge that they should have acquired only by interacting with the system. In order to counterbalanced the possible bias caused by allowing students to work in their time and location, they were asked to stick to the material provided and not to search in books or cooperate with other students. It is almost certain that the students worked as instructed[3].

6.2.1.4 Data gathering for building predictive models

The decision to allow students to work in as naturalistic situations as possible introduces technical and methodological problems. For example, it is difficult to record all aspects of students' interactions and record a talk-aloud protocol as it usually is the case in such studies. The technical problem of recording all aspects of student-system interaction was overcome by employing again the recording agent (also discussed in Section 3.3). However, it is still difficult to collect trustworthy students' reports.

First of all, in general, such reports require the externalisation of students' affective states which, due to lack of introspection skills, can be difficult for them. In addition, data collected from self-reports, similar to ones collected from questionnaires are often criticised because they can come from inaccurate expressions, especially for affective characteristics (O'Bryen, 1996; Stone et al., 1992) some of which are embarrassing to report or have negative connotations. Another problem is that self-report may influence students cognitive process while they are working but also the actual affective process that they are reporting (de Vicente, 2003; Masthoff and Gant, 2006). On the other hand, the potential of the method has been highlighted quite successfully a number of times. One example is de Vicente (2003) where also Spensley et al. (1990) and Issroff (1996) are cited. D'Mello et al. (2006b) also employ talk-alouds and collect data the analysis of which provides useful results. It seems therefore that the method can prove useful especially if some of the associated problems can be controlled.

Based on the above, and particularly in order to interfere as little as possible, it was decided to collect data during retrospective post-task walkthroughs where students

[3]Needless to say that, due to the unconventional requests, students who took part in the study were offered the opportunity of an additional tutorial if they did not grasp the material.

would watch replays of their interaction. These walkthroughs are similar to retrospective talk-aloud protocols described by Ericsson and Simon (1993). The difference here is that the aim is not to investigate the cognitive processes involved during the interaction but the walkthrough aims to record students' affective and emotional states. This part resembles the type of research conducted in D'Mello et al. (2006b) with the difference of the immediacy of the think-aloud procedure since it was deemed important that students interact in their own time, not influenced at all by external factors, observers or even the task of having to talk aloud.

While verbal reports have also been criticised (e.g., Nisbett and Wilson, 1977) many researchers provide positive results from carefully manipulated studies (for examples see Ericsson and Simon, 1993, pp.25). In addition, most reviews so far focus on studies that attempt to elicit verbalisations of cognitive processes. The effects of attempting to unravel affective processes are not so clear. Undoubtedly, challenges are also associated with this kind of report. First of all the time between the interaction and the walkthrough (Ericsson and Simon, 1993) could play an even more vital role than when investigating cognitive processes. For this reason an effort was made to conduct them as close as possible to when the student completed their interaction. Usually this was done the same or next day. The most prominent risk with this kind of report is that retrospective feelings can differ from feelings experienced (Masthoff and Gant, 2006; Ericsson and Simon, 1993). In addition, it is possible that students provide reports based on implicit theories behind the situation of how they should have felt or the effort they should have put in the task. In an attempt to mitigate these risks students were given a short questionnaire in advance (see Appendix C). The questionnaire had screen-shots of the system's pages and space for note-taking, which they were asked to complete after each interaction with the system. This methodology resembles 'confidence logs' or similar techniques of note-taking that are used in educational studies to record students' and tutors' interactions (Draper et al., 1994). The notes that students kept helped to stimulate their memory during the walkthrough. In addition, in order to keep within sensible limits the time that the walkthroughs would take these notes together with graphs based on their interaction (the technique described in Appendix E.2) served as a guide for choosing which particular interactions to target during the walkthroughs.

Finally, even if some of the students' reports are based on their own theories and not their actual states this is not necessarily problematic; at least at this stage where the main goal of the analysis is to investigate the machine learning methodology. The validity of course of the results is an issue that has to be researched further.

6.2.1.5 Situational factors and their values

The investigation in this part of the research covers a set of affective, emotional, motivational and other factors that relate to the situation and are encapsulated under one umbrella; situational factors (Porayska-Pomsta, 2003) as discussed at page 29 of Section 2.5.

The choices of factors and their possible values was not arbitrary. Their importance was inspired from other related literature (Porayska-Pomsta, 2003; Kort et al., 2001; Rozin and Cohen, 2003; D'Mello et al., 2006b; Conati, 2002; Andres et al., 2005) in similar contexts. Their possible relevance to the mode of interaction behind WALLIS was further identified during the pilot as well as the findings of the descriptive analysis of the ActiveMath project (see Section 2.5 and Porayska-Pomsta et al. (2008) for more details) where human tutors diagnose student's affective factors during an interaction with them. Consequently, the situational factors consist of :

- binary *emotional states*: *frustration*, *boredom*, *confusion* and a general category for *positive feelings* to represent satisfaction, joy, happiness etc.

  The reason behind the choice of the binary value of the factor was the fact that during the pilot it was identified that students find it easier to report the presence or absence of an emotion rather than its intensity. The general category, *positive feelings*, helped students, who often found it difficult to verbalise or differentiate their feelings, to be more expressive. D'Mello et al. (2006b) report similar findings when trying to use *eureka* as a state for signifying happiness or generally an expression of positive affect caused from answering correctly. Instead of 'eureka' or 'positive affect' which was attempted during the pilot study, it was established that the term 'positive feeling' helped students articulate their emotions more easily and minised discussions, unnecessary clarifications and confusions about the term and its use.

- relative change factors: *confidence*, *interest* and *effort*.

  During the pilot, students found it easier to report factor changes rather than choosing a specific value from a range (such as low, medium, high) which, for them, do not necessarily have a meaning. It was decided therefore to use relative change values (such as increase, decrease) since the cases where the factor changes are the ones that are of particular interest to this research. In addition, more instances of the same event become available rather than specific changes from low to medium or from medium to high. This increases the power of any statistical analysis and facilitates any machine learning technique to come up with more general results.

- factors that relate to and depend on the situation: *difficulty* of the current material or step, *time spent* measured in standard deviations from the time all students took to complete the same task, *correctness* of an answer, *interaction type* (input box, radio button etc.)

- and *student characteristics* (previous performance in related skills and post-test marks, awareness of the material as well as the probability they need help in order to answer correctly, and whether the interaction at that point was beneficial according to the model described in Chapter 5.3).

6.2.1.6 Process

During the replays students were asked to comment on the situational factors mentioned above and although a certain structure was followed the discussion was quite open. In this way, it was possible to remind students to express their emotions when they forgot, but also to pause the replay at particular moments indicated by the student from their notes or where an overt behaviour was observed in advance (based on graphs of their interaction and replays which were seen in advance). For every state expressed, with the means of a mechanism during the replay, a timestamp was associated with the log file. It is worth pointing out that participants were strongly encouraged to express anything in relation to affect or motivation that was not included in the description of the task. In the pilot study this was proven helpful as additional factors were added. During this study, 3 additional emotions were reported by students themselves (*happiness*, *satisfaction* and *upset*) which were included as occurrences under the positive feeling and frustration factors, respectively. In addition, there were even some cases where students reported a combination of affective states (particularly confusion with frustration and interest with satisfaction). These were included in the data and were treated as two different instances with the same timestamps.

Apart from recording the factor value and student comments the other goal of the walkthrough was to gain extra insight in order to guide any subsequent data analysis. In particular, it was important to validate previous findings from pilot studies, that there would be no or just few occurrences caused by characteristics of the system (e.g., frustration from delays or confusions because of the buttons). In addition, during the talk-aloud, the open structure of the discussion meant that the the source behind the reported affect could be disambiguated and, in cases where this was not clear, the immediate preceding action of the student could be identified. It should be noted here that in order to contain the task, the goal of the walkthroughs was not to perform an in-depth qualitative analysis but to investigate the possibility of the methodology reported yielding useful observations for machine learning analysis. One of the most important findings of this stage (apart from the expressed factors) was that it became evident that, in most cases, the students reports were not caused by some local, immediately preceding action but because of actions and feedback of the system, and student reactions, that had been building up for quite a while. This is further discussed in Section 6.2.3.

6.2.2 Descriptive analysis

A descriptive analysis provided indications for the general patterns and guided the machine learning analysis, which is reported in the next section. In order to analyse the results, the timestamps from the log files were matched against the immediately preceding student-system interactions. Note that the environment is constrained and therefore the student actions have a one-to-one mapping to a system re-action which

student action	action characteristics	system re-action
provide answer	partially correct	corrective feedback
	incorrect	negative feedback
	correct	positive feedback
help request	answer partial	hint based on error (corrective)
	1st ... penultimate hint	increasingly revealing hints
	last hint	very descriptive hint
	solution	full solution
suggestion request	while on page	explanation about page's goal
	after completing goals	further page suggestion
exit attempt	stay on page go to example exit to other page	prompt to select subaction (see Section 5.4.3.2)
other	select page	page delivery
	reading theory,example	none explicitly
	reading exercise	none explicitly

Table 6.1: Student actions and their mapping to system reactions

is deterministic. The possible student-system interactions are described in Table 6.1. For example a wrong answer is bound to generate negative feedback, a partially correct answer will generate corrective feedback, and so on. These interactions are often spontaneous and it is hard to differentiate whether the students' report refers to the system's reaction or their own action. Therefore, the analysis is conducted with the assumption that a report refers to either the student's or system's immediately preceding action.

student action	system re-action	percentage						
		emotional states				other factors		
		frustration	boredom	confusion	positive feelings	confidence	interest	effort
provide answer	corrective feedback	12.68	0	13.43	13.43	15.22	6.56	25.3
	negative feedback	22.54	0	18.98	18.98	20.42	0	18.07
	positive feedback‘	0	12.82	0	0	0.69	0	16.27
help request	solution	7.04	20.51	4.17	4.17	2.08	14.75	3.01
	hint_corrective	12.68	0	2.31	2.31	16.26	3.28	4.82
	hint_low	19.72	17.95	5.09	5.09	4.15	4.92	1.81
	hint_high	9.86	2.56	13.43	13.43	13.84	8.2	2.41
suggestion request	inpage	2.82	0	0	0	4.5	0	1.81
	exit prompt	7.04	0	0	0	0	11.48	12.65
	end of page	0	0	0	0	3.11	19.67	4.22
other	selecting pages	0	30.77	14.35	14.35	0.69	13.11	1.81
	reading theory,example	1.41	12.82	22.69	22.69	10.03	14.75	2.41
	reading exercise	4.23	2.56	5.56	5.56	9	3.28	5.42
	overall	7.64	4.2	23.25	9.36	31.11	6.57	17.87

Table 6.2: Percentages of observation with situational factors

Table 6.2 presents the percentage of situational factor reported in relation to the immediately preceding student-system interaction as well as their overall frequency. For example, 12.68% of frustration reports were after a partially correct answer, 18.98% of confusion reports were after the provided a wrong answer (and therefore received negative feedback from the system). From the table, the difference between some factors is notable. Factors *confidence*, *effort* were the easiest for students to report. In terms of the *emotional* factor *confusion* was the most reported value. Similar results have been reported in other studies (de Vicente, 2003; D'Mello et al., 2006b). Particularly (de Vicente, 2003, p.55) observes that *confidence* and *effort* were also the most reported factors in students' self-reports.

However, a descriptive analysis, cannot differentiate the reasons behind some of the reported factors. Therefore, its purpose here was limited to help towards the secondary goal of this part of the research: to identify which emotional and motivational aspects appear during the interaction with a system like WALLIS. Similar studies (for example the one mentioned at the Background in the context of the ActiveMath project - see page 28) have highlighted the importance of the factors included in this study but usually from a tutor-perspective. Since the current study investigated the issue from a student-perspective, the results are suitably valid in the sense that they reflect students' concerns. Despite the fact that some of the reported factors may not be relevant for guiding the adaptation and feedback further, they are useful in terms of deciding which factors students think are important during their interaction.

Based on the descriptive analysis, *confidence* and *effort* were considered in depth for the rest of the analysis. The *emotional state* factor, despite the limited amount of data, is also discussed in some more detail than the rest of the factors.

6.2.3 Machine learning analysis

As explained before, the primary goal of this part of the research is to investigate the possibility of automatically predicting affective states based on students' actions. What became apparent during the walkthroughs and the data investigation process is that in order to have more accurate predictions, a broader scope of historical information is needed to determine what really causes the reported state of the factor. Many similar actions in the data produce completely different reports of the emotional or other situational factors. Most of the students when verbalising about a factor were often referring back to previous actions (for example *"Here my confidence decreased. It is the third time I try and I get it wrong despite it's [the system] being helping me."*). In order to take such statements into account it seems that the history of actions needs to be included in the prediction task. There are other cases where this is apparent, particularly when the values of factors do not change. These cases are not directly observed from students' comments. A good example is the hint-abuse behaviour mentioned in Chapter 4. The effort of students who do not read the hints carefully but are just requesting hints to reach a solution is probably different if compared to the students who spend some time to interpret the given hints. However, there is a lack of report of factor value changes under such conditions. If historical information is not taken into account, hint requests appear directly associated with increased effort. This is not necessarily true. Similar examples apply for confidence. Asking for a hint after attempting to answer once is not the same as asking for a hint after having tried to answer several times. D'Mello et al. (2006b) report a similar problem in relation to boredom detection. Other relevant research (de Vicente, 2003; Forbes-Riley and Litman, 2004) as well as the study with human tutors reported in the next section (see Section 6.6.4) confirm this further.

In order to perform this data analysis, for each factor to be predicted, vectors (instances) are presented to Weka's J4.8 algorithm (the algorithm is described in Section 3.4.1 in more detail). These vectors are constructed from the log files by an ad-hoc extraction algorithm that performs the following actions:

1. matches the timestamps of the expressed situational factors with the immediately preceding student-system action

2. adds the `action` and the relevant `situational factor` values as elements of the vector

    - *correctness* of student action (correct, incorrect, partially correct[4]).

[4] according to a list of common misconceptions

- the *time* that the step took overall (discretised according to the method described in Section 3.4.6.2).
- students' *previous performance* (a numeric value from 1 to 5).
- *difficulty* of the item under question (a pre-determined numeric value from 1 to 5).
- according to the models in Chapter 5.3, the prediction for how beneficial the interaction (*b*eneflnt) for this skill was when help was requested, and the probability that it was needed at that point (*need4help*).

3. It extracts the history before the action. The vector also includes the history of the actions. This is represented as a vector whose elements encode the number of times each type of possible action (e.g., a hint request) occurred in a relevant time window. The time window spans back to the last change of the factor under investigation, or to the start of the relevant exercise or situation (e.g., reading an example, solving the exercise) — whichever is sooner.

4. includes the nominal class for prediction as the last element. For *confidence*, *interest*, and *effort* the class takes values that depict relative or extreme changes: *decrease*, *increase* and *extr_decrease*, *extr_increase*. For the emotional state factor the values are the different emotions described in section 6.2.1.5.

In order to help the tree induction algorithm generate satisfactory results, the data are preprocessed as follws:

- In order to align the data accordingly and because not all situational factors are pertinent in every situation (e.g., correctness is not relevant when reading theory) a merging of some variables was necessary to avoid having missing values for some of the instances. That is, if the vector was of the form:

  ```
  [action,...,correctness, need4help, ...]
  ```

  it would have resulted in missing values for some of the instances where the correctness or the probability of need for help are not applicable. Therefore, when appropriate (i.e., in the case of the action being an answer) the `action` string is conjoined with the `correctness` resulting in the following values.

`confirm_answer_{correct,partial,incorrect}`

Similarly, when the action is related to the student asking question, the `need4hint` variable is added so as to include the additional prediction of the student requiring the hint at that point or not. This results in action types such as

`hint_info_{yes,no}`

- depending on the prediction task, a vector of different features is chosen using Fast Correlation Based Filtering (discussed in Section 3.4.1)

Running the algorithm on the set of data (henceforth referred to as `Std_Set_A`) derived from all the above processing results in biased rules that are influenced by the fact that the set only includes factor changes rather than additional situations where nothing changed. Therefore, for each factor, instances are constructed with data that involve the same student and system actions but after which no factor value was expressed by the student (these will be referred to as `Std_Set_B`). The class in that case is `no_report`. Whenever possible, the merged data are presented to the learning algorithm. However, in some cases (e.g., for the emotional factor) the instances where there was no report are much more frequent than the cases where there was a report. The tree induction algorithm prefers splitting data based on attributes with the highest information gain, and therefore tends to misclassify the instances where there was a report in relation to a factor. In these cases only `Std_Set_A` is used but the results are treated with more caution, providing hypotheses for further testing in the future to establish the exact circumstances for the changes.

Because of the limited size and sparsity of the data it is difficult to come up with a representative set that could be used for evaluation. Therefore, stratified cross-validation is employed with a fixed number of folds for measuring the performance of the tree. As an indication of the quality of the tree, apart from the percentage of correct classifications the Kappa statistic is consulted and the F-ratios (Witten and Frank, 2005).

6.2.4 Results

6.2.4.1 Confidence

In total there were 289 reports in relation to confidence, these comprise `conf_Set_A`. Another 249 instances were extracted from the data that involve the same events reported in the first set but for which no factor value was expressed by the student (these comprise `conf_Set_B`) . Figure 6.3 shows a graphical representation of part of the tree resulting from the merged sets of instances with attributes *action*, *difficulty*, student previous *performance*, last *answer*, *time* between the last two actions, and the *history* vector. 90.9% of the cases are classified correctly and Kappa is 0.87 which can be considered a satisfactory result.

As an example, the rules associated with the *confirm answer* action of the student are described here. The tree suggests that when students confirm an incorrect answer (node `a`) having previously requested at least one hint, their confidence decreases (leaf `b`). If no hints were requested and if they previously had many (more than 2) incorrect answers then they report that their confidence decreases (the two misclassified instances in leaf `c` are of an extreme decrease). Otherwise it seems that the outcome depends on students' previous knowledge, the difficulty of the question and the necessity of the help request (node `d`). In particular, node `e` of the tree suggests that students with high previous knowledge and for whom the bayesian model predicts that they need help reported that their confidence increased, whereas the rest did not provide a report. Node `f` suggests that students with low previous knowledge, who did not ask for hints, reported that their confidence was decreased for questions after making few incorrect attempts to answer.

Similarly we can interpret the rest of the tree. Notice the rule associated with node (`g`) where students are requesting hints after submitting a wrong answer. Their reports depend on having previously replied partially correctly. Not surprisingly, it seems that the fact that students who were previously approaching the solution and now receive negative feedback decreases their confidence, especially of the ones who have not spent enough time to read the hints carefully. The others probably get over this problem by spending more time hopefully reading the feedback and trying to understand.

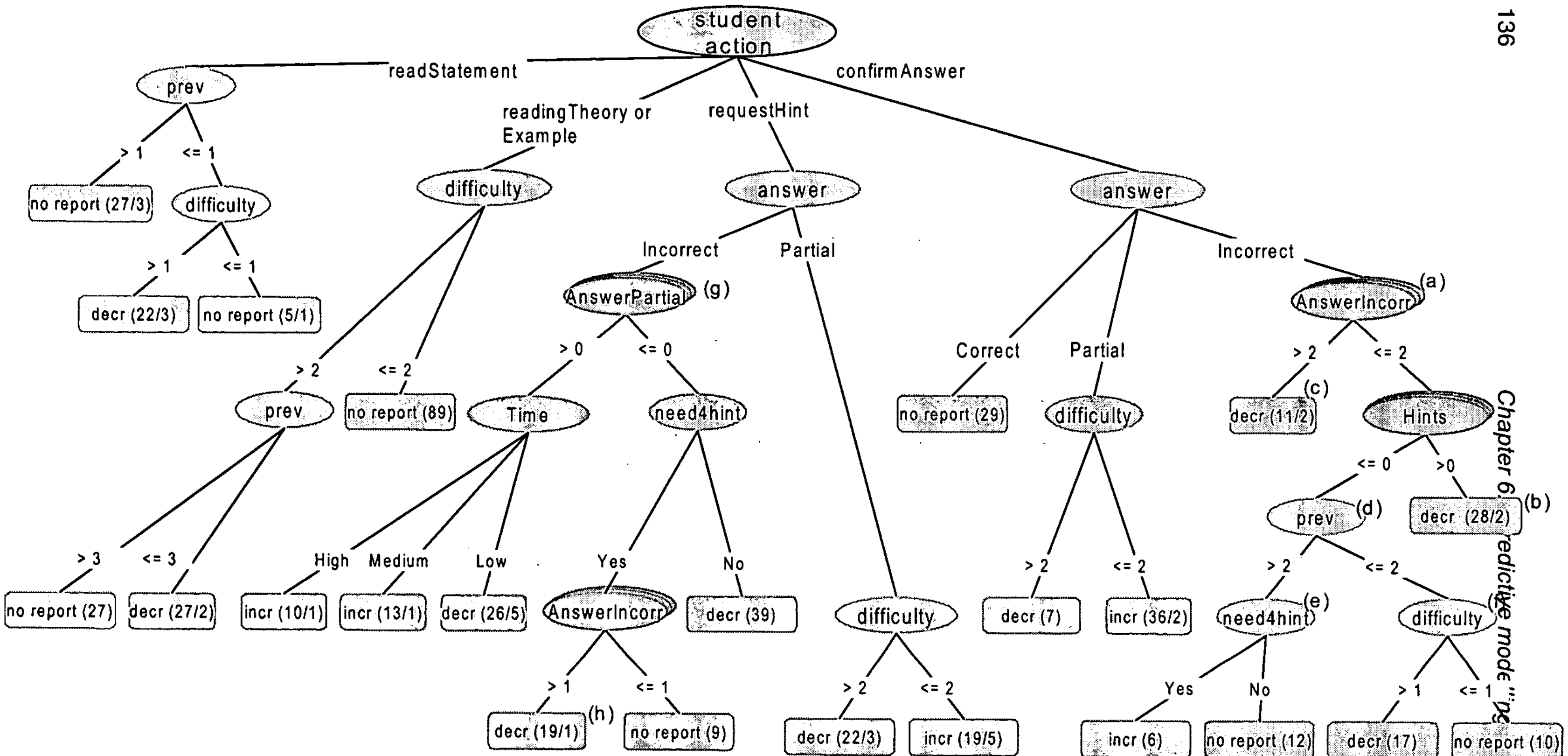


Figure 6.1: Graphical representation of the decision tree for confidence. Each node represents an attribute and the labels on the edges between nodes indicate the possible values of the parent attribute. Following a path from the root to a leaf creates a rule that shows the value of the factor over the values of the attributes in the path. The numbers next to the designated class show the number of instances correctly classified by this rule over the misclassified ones. Nodes with layers (e.g., node a) originate from the vector of historical attributes. For example `AnswerIncorr > 2` shows that in the relevant time window students provided more than two incorrect answers. The labeled nodes and leafs (a-h) indicate examples discussed on page 133.

The latter brings to mind the results from the investigation on hint reflection and learning in Section 4.5.3.2. Note that the same pattern is associated with increased effort in the effort tree, presented in the next section. The remaining rules seem, in general, equally intuitively plausible. However, it would be hard to generate all these rules intuitively. In addition, in some cases (especially where misclassifications occur), detailed analysis of the raw data has to be conducted in order to establish the reasons behind them. Although time consuming, this process generates hypotheses for further studies.

At this stage, it is worth describing a common problem. The rule ending with leaf h indicates that, when requesting a hint for an incorrect answer and if the learner had replied incorrectly more than once in the past but no partially correct answers were given, then their confidence decreases. It seems that the negative feedback that WaLLiS provides without any corrections (due to the difficulty sometimes in identifying students' error) could dent students' confidence. As an aside, it is worth noting here that, when this happened, around 63% of the students attempted to quit the page. Therefore, once again, the importance of adapting the environment according to correctly diagnosed students' affective states is important. In this case the system could intervene appropriately and attempt to boost student's confidence despite the fact that the exact error is not traceable. The other part of this rule is also interesting. There are no reports for the cases where there was only one (or none) incorrect answers before. This raises a question. In most of the cases where this happens, the reason is that either the student commented on another factor or that simply there was no change to the factor. Therefore the value `no_report` can be interpreted as no change. However, as will be further explained in the Section 6.3, there are also cases where the absence of reports was due to the lack of metacognitive skills of the student and difficulties when they were prompted to express their affective state.

Nevertheless, despite the fact that this microanalysis is time consuming, it can help improve the quality of the diagnostic process of the system and consequently make it more effective.

6.2.4.2 Effort

Effort was a more complicated factor for the students to report. It was linked both to boredom and interest, and during the walkthroughs it was often difficult to separate them. For example, many students reported a decrease in effort due to the lack of confidence to answer the question. Because of that, they relied on the system to give them the answer but some admitted that with a bit more effort they could have answered the question. In addition, since a decrease in effort has negative connotation it was one of the factors for which not only decrease was less reported but there were many cases where the lack of reports is more likely to be associated with a decrease in effort than in a genuine no change. There were 213 reports in relation to effort, these comprise `effort_Set_A`. Another 153 instances were in `effort_Set_B` with data that involve the same events but with no factor value expressed by the student. Figure 6.2 shows a graphic representation of part of the tree from `effort_Set_AB`. 89.159% of the cases are classified correctly and $Kappa = 0.799$ which is a satisfactory result. As an example, it is worth describing the rules associated with the *hint request* action of the student. The tree derived suggests that the effort of students who request a hint when they have answered incorrectly more than once, depends on the difficulty of the question (see branch starting at node a of Figure 6.2). There is a lack of reports to be able to deduce doubtful rules, but it is clear that when the item is very easy then students reported that it was their reduced effort which made them ask for more hints from the system. The same applies for slightly more difficult questions but then it seems that time plays a role (branch b). When students do not spend enough time in the step it seems that they report putting lower effort in the task than when they spend more time on it. Although there are only four instances to justify the latter rule, it seems intuitive enough and further studies should investigate this in more detail.

It is interesting to observe that in this case the variable `need4help` and `benefInter` are included in more rules than that of confidence. For example, branch (c) suggests that when student did not have many previous attempts in answering the question there was generally no report from them in terms of their effort, apart from the cases where the Bayesian model for *need for help* predicts that they need help indeed. In 6 instances it was reported that their effort was increased. These appear in leaves (d) and (e).

In particular, in leaf e they appear as misclassifications due to the lack of reports. Once again, the amount of data are not enough to derive unequivocal rules but these are enough to construct hypotheses for future work. The above also demonstrate the potential of measurement such as the *need for help* which are not directly observable but derived from the rest of the students' interactions. Inducing a tree with the same set of data without this variable would result in more misclassifications.

As mentioned at the beginning of this section. In the case of effort it is more likely that the lack of reports is associated with a decrease rather than a genuine no change. For example, under branch (f), the effort of students who do not spend much time with the hints carefully is more likely to signify a decrease (as indicated by the 3 misclassified cases) rather that a lack of change. Based on the tenet of this thesis not to have intuitive results but based on data it seems that more data need to be collected in order to test this kind of hypotheses. On the one side, the full tree at Appendix D provides more hypotheses for future work. On the other side, the issue of effort having negative connotation needs to be dealt with appropriately and other ways should be devised to elicit rules about effort. A peer observation method, also mentioned in Section 6.5, is a possible approach. Other possibilities are discussed at Chapter 7.

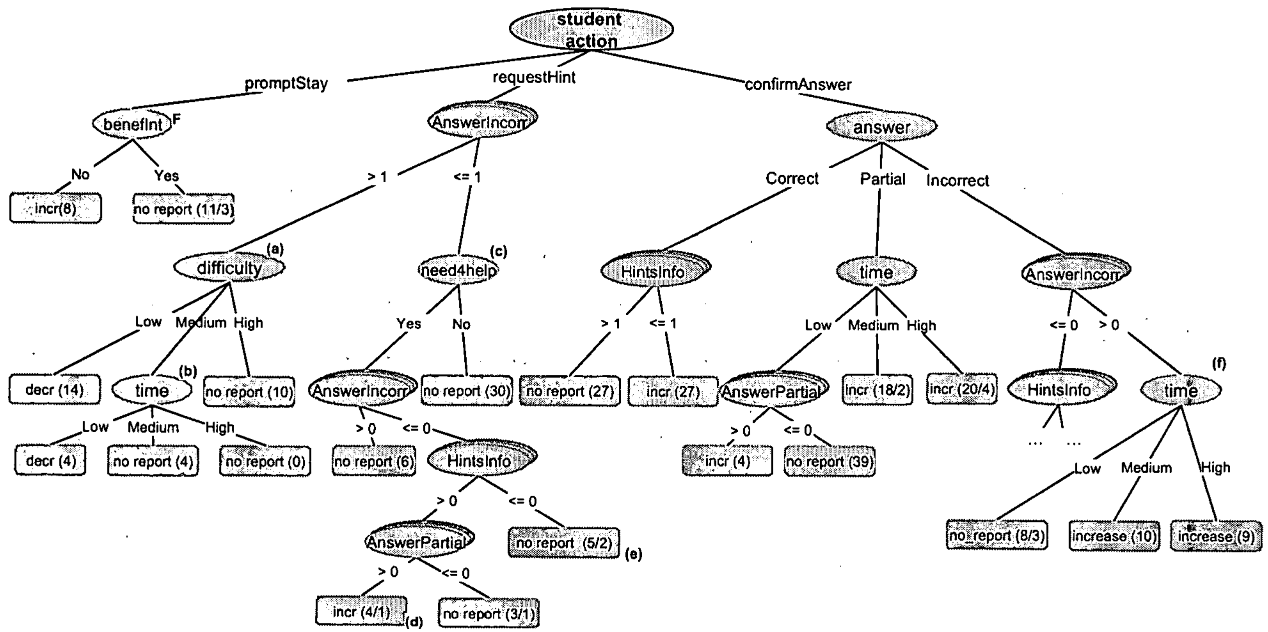


Figure 6.2: Excerpt from the decision tree for `effort` trained on `effort_Set_AB`. The labeled nodes and leafs (a-f) mark examples discussed on page 6.2.4.2.

6.2.4.3 Emotional state

As Section 6.2.2 shows, the emotional state factor was mentioned quite often. However, the sparsity of the data impedes the machine learning analysis. This is because the majority of the events which are associated with a reported emotion (e.g., confusion) occur more times without any associated report and, as discussed in Section 6.2.3, the machine learning algorithm prefers to misclassify the instances where there is a report. If only the set of instances with reports is considered, then degenerate rules are produced. For example, a rule could predict that if students are answering correctly they report that they are confused. This is obviously a rule that cannot be used as such. Once again, there is nothing to contradict this rule to generate a more elaborate rule.

As explained before, the walkthroughs were not designed for a rich qualitative analysis as this was beyond the scope of this research. However, it is worth mentioning that repetitive incorrect answers caused confusion or frustration with no clear distinction between them. For some students it seemed that frustration was often caused because of their confusion especially when they could not immediately interpret the hint that they were receiving. Frustration was also associated with the first levels of hints that did not necessarily address students' needs immediately. Positive feelings were also reported. Perhaps not surprisingly, they were particularly associated with correct answers but also with corrective hints and after solution requests. From the walkthroughs it is more evident that this was the case when students understood the feedback provided and felt satisfied to progress. Additional positive states were reported after suggestion from the system on which pages to select and what to do.

6.2.4.4 Other factors

As Table 6.2 shows a very limited amount of data was collected for the rest of the factors. In addition, for some of them, there are lots of cases where there are conflicting evidence (i.e. events where a change was expressed with equal or more events which had no report associated with). This makes the statistical and machine learning analysis unfeasible. Nevertheless, the data collected are interesting for further qualitative analysis and for generating hypothesis. Some of the hypothesis are provided below for the benefit of future research.

- Requesting hints is directly related to confusion. Often the request of consequiteve hints indicates confused students who want to read all the hints together in order to make sense. They reported that they treat the bottom-out hint as an example. The ones who manifest more interest and effort attempt the exercise again.

- Requesting for hints quickly has to be evaluated in the greatest context of a student's interaction. Some students that have interacted successfully with the exercise in the past are going through the exercise quickly by requesting hints not necessarily because they are bored or because of their lack of effort but because they want to confirm their understanding. The measurement of *beneficial interaction* devised in the previous chapter could prove useful in predicting these cases.

- Interest seems to be related to students' accessing theory pages rather than just the examples and following the links suggested by the system. In addition, higher interest was also reported when students returned to an exercise even though they completed previously. It should be noted that decreased interest was rarely reported, probably because of its negative connotation.

- Students reported positive feelings in relation to positive feedback received from the system. In particular, these are more evident when their confidence in their answer was low. It seems therefore, that there is a scope in including factors together when performing the data mining.

6.3 Discussion on Predictive Modelling From Student Perspective

The methodology described in the previous section and the study conducted yielded useful results that contribute to the overall task of detecting students' affective states. Through the walkthroughs it was established that the commonly reported factors in the literature appear also in the context of this research. They also provided a fundamental insight which influenced the machine learning analysis. The history of the interaction is very decisive in determining changes in students' affective states. Finally, despite the limited amount of data, two of the situational factors *confidence* and *effort* were reported sufficiently enough times in order to perform a machine learning analysis.

The machine learning analysis and particularly rule induction provides a mechanism for deriving a predictive model of affective states in the form of rules that are based on student actions. For every factor an inspectable tree is derived which can form part of a predictive model. The investigation so far highlights some important issues for further work in relation to the data as well as the methodology behind their collection and analysis.

As discussed in the last section, there is a challenge introduced by the assumption that 'no report' is the same as 'no change'. This is not always true and is something that was not anticipated in advance of the data collection. In the case of this study, due to the way the walkthroughs were conducted, students were prompted in specific places (based on replaying their interaction in advance and from their notes and analysis performed on the data from the pilot study). However, this is rather subjective and prone to introduce bias. For some of the factors, the sparsity of data seems not to affect our results. Still, the new patterns that have arisen have to be revisited in future studies and a more objective way needs to be devised that could prompt students in particular places where a pattern seems to arise. This is an issue that is not specific only to the current study but one that should be carefully reviewed in other similar studies particularly when attempting to analyse data automatically rather than employing qualitative techniques.

It is interesting to observe that 91.54% of the patterns which are not associated with any report, are from students with low or very low previous knowledge of the material and low post test scores. In addition, the majority of 'missing' values is for the confusion or interest factor. This confirms the intuitive fact that lower ability students are not always insightful about their emotions (see also Aleven et al., 2003) and hence they report some factors (especially confusion) less accurately. It is clear therefore that the metacognitive abilities of the participants play an important role and is something that needs to be thought through carefully in advance (for example, it would have been interesting to test for metacognitive skills when selecting subjects).

A related issue is the dependency of the analysis and results on these self reports. The limitations of the approach were discussed in Section 6.2.1.4. From the study, it seems that the problems associated with the retrospective walkthroughs do not impede the machine learning analysis of the *confidence* factor that does not have such a negative connotation as *effort*. Students may be "embarrassed" to report values that will make them look lazy (like low effort or boredom) but are keen to report that they are exerting effort. Therefore, students seem a more valid source of evidence for reporting their own emotions and factors like confidence but in some cases they may be reporting biased from their own theories about what a good student should be feeling or how much effort they should be exerting. In these cases it seems that tutors are a more suitable source for the way boredom or effort is perceived as they are also the ones who determine what the effects of this diagnosis are.

Finally, it is worth pointing out that although the walkthroughs were not designed to elicit qualitative results, they still contain knowledge that was not analysed in this thesis since the main goal was to investigate the machine analysis methodology. In the future it would be interesting to improve the quality of the results taking into account different sources of information (what is usually referred to as *triangulation*).

6.4 Predictive modelling from tutor perspective

The previous discussion highlights the limitations in relation to predictive modelling based on data from student self-reports. One of the aspects that was clearly identified is students' tendency to provide more accurate reports on states that have positive connotations. The lack of introspection skills as well as the general issues with self-report introduce bias in the model. It seems therefore natural to attempt to remove this bias by introducing a tutor perspective into the diagnosis. While the student perspective is still useful, it is expected that the way tutors perceive at least some of the situational factors, is equally important. In the end, it is their inferences in relation to learners' affective states (together with inferences about cognitive states of course) that are the main driving force behind the interaction (Porayska-Pomsta et al., to appear). On the other hand, as discussed in Section 2.3, tutors are not necessarily good at detecting and responding to affective states. Therefore, both tutors' and students' perspectives seem important and need to be taken into account.

Towards this goal, an attempt to derive tutors' perspectives from watching recorded student-system interactions was made. The pilot study, raised some issues and difficulties, suggesting that a full study was not feasible at this stage. Section 6.5 describes these difficulties as well as insights for future work.

In order to investigate the issue of tutor perspective, a dataset collected in the context of the EU-funded project LeActiveMath (LeAM, 2003) is well suited. In particular, the dataset consists of tutors' affective diagnosis collected from empirical studies with tutors and students interacting with each other in an environment approximating an ILE. The study and the analysis is described in Section 6.6[5]

While the specific goals and focus of the aforementioned study are a bit different compared to the ones of this thesis, the data collected are particularly useful and can be used to investigate how the machine learning methodology which was employed for analysing students' perspective can also be used to analyse more complex data. These data consist of dialogues and not of constrained student actions as when students were interacting with WALLIS. The subsections of Section 6.6 describe the steps that

[5]The author of this thesis was involved as a developer in the project and collaborated with the researchers of this particular study.

needed to be done in order to analyse this dataset as well as other issues that emerged. As outlined before, more general issues emerging from this as well as the previous studies are discussed in the next Chapter.

6.5 Employing replays of students' interactions

In order to investigate the possibility of conducting a study to collect affective diagnosis data from tutor perspective a pilot was performed with recorded students' interactions with the system from the `GIC04` dataset.

Two experienced tutors of mathematics were asked to participate in the study. The choice of these tutors was based on the fact that they had some experience with WALLIS (since they were both teaching the `MM` course) and are quite positive about the idea of computer-based teaching and learning. It was considered important to select tutors who are aware of the system and supportive of its use. It was hoped that this would ease the task for them.

The tutors were requested, similarly to the students at the study before, to observe the interaction and to comment on the affective state of the student during the interaction. This part of the study was inspired by a similar motivation diagnosis study of de Vicente (2003) (described briefly in Section 2.5). The difference here is that the main goal is to investigate the use of machine learning techniques to analyse the data rather than the qualititative analysis conducted in de Vicente's work.

The pilot revealed the following issues:

- The tutors needed more information about the background of the student. While this was rectified during the pilot by replaying a lengthier time window of interaction, both tutors often mentioned that knowing the ability of the students prior to their interaction would help them in certain cases. This was particularly when commenting on students' effort in relation to the requests for help. In addition, it was interesting that the tutors requested more information about the prerequisites that the students have accessed as well as the overall goals of their course and their interaction with the system.

- Given the different medium, it was difficult for tutors to say when reading a page or an example is enough and if studying the content in this page and the examples would have been enough for the students. During early discussions, it emerged that using the models derived in Chapter 5 of the beneficial interaction and the need for help would be useful. While these were provided to the tutors they did not help in eliciting enough data.
- The tutors were not cued in to all aspects of student interaction. There were many cases where one could hypothesise about some of the affective characteristics of the students but the tutors consistently reported that it was very difficult for them to monitor and derive useful results from all aspects of students' interactions. The initial expectation that involving tutors who are familiar with the system did not seem to help.
- Despite the replay tool's capabilities to slow down and replay interactions the tutors were often lost as to what exactly the student was doing.
- Even when verbalisations seemed to emerge, the tutors were in constant conflict about their comment, often second-guessing their own choices.

Although some of these difficulties could have been rectified in an actual study (for example, more background about the students could be provided to the tutors in advance) some others were quite discouraging. Both tutors commented that they found it particularly difficult to comment on aspects of students' behaviour mostly because they were not involved in the situation as it occurred. It was hard for them to imagine the whole situation and how students would have felt and what the goals were behind a student's actions. Perhaps their own initial expectations were coming into play. Even before the study, the two tutors were very sceptical about their ability to derive any useful information about the students by only watching their interaction. Another difficulty is that of the tutors' involvement in the situation. Both the tutors explicitly mentioned that the diagnosis task would be easier, more natural, if they were interacting directly with the students.

de Vicente (2003) reports quite different results. Although the tutors were also initially quite convinced that it it would be 'virtually impossible for them to extract any useful information' (de Vicente, 2003, p95) about students' motivational states without being able to see them, the results were quite satisfactory. Looking into the difference of the two studies it seems that several reasons could play a role for the difficulties in the current study.

First of all, the context of the studies, as well as the two systems, are substantially different. The interaction with WALLIS takes from half to one hour, requires more complex behaviour from students and, as explained before, are under realistic conditions. de Vicente employs a simple mock environment for language learning that contains only multiple choice questions. Therefore, although the methodology appears quite promising it may not be scalalble.

In addition, de Vicente mentions that due to technical difficulties the mouse movements where not fully replayed and that having the ability to replay mouse movements could have helped tutors derive more comments. It was inspired by this, that the interaction capture agent of WALLIS records all mouse movements. However, it seems that sometimes this hindered rather than facilitated the knowledge elicitation process as it was originally thought. The tutors were often misguided by the mouse movements and attempted to interpret them in vain as they were often quite ambiguous. For example, a student could click on an answer box in passing with the mouse and then continue typing there while the mouse could be in a corner of the page so as not to obstruct the answering process. The second tutor, in particular, who paid more attention to the mouse movements, got confused by situations like this. Although it was possible to pause and also slow down the replay this brought unnecessary discussions and confused the tutor further as to the exact amount of time the student spend in the step, thus further hindering the elicitation process.

Finally, the tutors in de Vicente's study were postgraduate students. One could hypothesise that perhaps they are closer to and can empathise with the feelings of other students better than the turos in the current pilot. The task (learning a foreign language) is also something that perhaps the participants could see themselves taking part in. Therefore, they may have drawn on their own experiences as students for their comments. In addition, it would be interesting to know as well if some of them have taken courses that employ computers for learning. Even their general computer literacy or familirisation with the research field could play a role here (the volunteers where postgraduate students in the School of Informatics). However, the participants in the current pilot were already experienced mathematicians. This raises the issue of the appropriateness of the participants in studies like this. The above suggest that it may be a good idea to use other students (and particularly peers who would have interacted with the same system) as another source of diagnosis. On the other hand, perhaps a combination of HCI experts, familiar with both ILEs and educational aspects, would prove more appropriate. After all, the issue investigated is not so closely related to the exact domain of the concept taught to need domain expertise. A combination of skills is needed to be able to accomplish all that this task demands and the issue of who is 'expert' for this kind of knowledge elicitation is not an easy one to answer.

The aforementioned problems, the lack of volunteers, budget, time and other practical considerations rendered the full study unfeasible. While it was possible to rectify some of the above problems, particularly the issue of choosing the right experts was quite restricting. In addition, the issue of the involvement of tutors in a realistic interactions rather than a replay suggested that perhaps a different kind of study would be more appropriate. A possible approach often used in the field of AIEd is what is often referred to as 'wizard of Oz' experiments. In this experiment students could interact with WALLIS but the feedback provided, suggestions etc would be (perhaps partially) provided by a human.

Although this remains as further work, the following section describes the study conducted in the context of an EU-funded project; LeActiveMath (LeAM, 2003) and addresses some of the issues discussed above.

6.6 Employing tutor-student interactions

The previous sections detailed the problems with student reports (mainly the bias introduced for the factors that have negative connotation) and the last section the problems with the pilot where tutors watched students' interactions (mainly the fact that they felt they should be more involved during students' interaction). This section employs a machine learning analysis, similar to the one employed for analysing students' perspective, for a dataset that consists of tutors' affective diagnosis collected from empirical studies with tutors and students interacting with each other in an environment approximating an ITS.

The following subsection describes the context and overall goals of this study and how they fit in the context of this thesis. Subsection 6.6.2 presents the data collection methodology that resulted to data that consist of dialogues between tutor and students and synchronised tutors' verbalisations of students affective states. In order to be able to machine analyse these data, the communicative goals behind the dialogues needed to be coded in order to represent the different types of actions on behalf of the student. The methodology and types derived are presented in Section 6.6.3. The machine learning analysis is described in Section 6.6.4. Although the methodology is similar to the one behind analysing the data from student perspective, the nature of the data imposes some differences which are also described there. Finally, Section 6.6.5 presents the results of the analysis and Section 6.7 highlights some of the issues associated with this study.

6.6.1 The context and goals of the study

The machine learning analysis described below was performed on a set of data collected from empirical studies for the EU-funded project LeActiveMath (LeAM, 2003). The general goals of the project were multifold. Here only the related goals are briefly presented (for more details see LeAM, 2003; Andres et al., 2005a; Porayska-Pomsta et al., 2008). As described in one of the reports describing the deliverable of the project (Andres et al., 2005a) the overall aim of the empirical studies the data of which are analysed below were (a) to collect natural language dialogues for the purpose of in-forming the design of the dialogue manager component of LeActiveMath and (b) to provide relevant information as to the affective, motivational and cognitive factors that impact tutors' decisions and students' learning. More specifically, as described in (Andres et al., 2005a; Porayska-Pomsta et al., 2008), the research aimed to establish:

- which situational factors are important when tutors engage in a dialogue with students in the domain of mathematics
- the extent to which human tutors are able to identify situational factors, and which affective states and other situational factors they diagnose
- what sources of evidence (cues) tutors rely on in making their diagnoses and diagnosing different situational factors.

The overall goals therefore of the above study are close to the context of the aims of this thesis. On the one hand, the study facilitated the identification of factors that play a role in this domain and context and influenced the design of the study with the students in Section 6.4. Some of the influencing results were overviewed on page 28 of the Background chapter. On the other hand, the data collected can contribute to the prediction of student affective and emotive states based on their actions.

6.6.2 Procedure and data collection methodology

The student-tutor interaction was through a chat interface. This is presented in detail in Appendix C.2.1. The tutors' task consisted of tutoring the student on the chain rule, using exercises that they saw fit. Students were told that they would receive online instruction, relevant to their current mathematics course, from a human tutor. Tutors and students were trained to use their respective interfaces, prior to interaction and were told that this would be their only means of communication.

In addition, tutors were asked to talk aloud about any aspect of the interaction while they engaged in tutoring. In addition, they were instructed to select situational factors, using a *factor selection tool*, every time they provided the student with feedback. They could do this either by clicking a "submit", or a "no change" button to indicate explicitly that there was no change in the factors between the previous and the current situation.

A set of pre-defined factors was provided to the tutors. These were not mandatory as the tutors were not required to select any of them at any point. The tutors were given an opportunity to specify situations in their own terms by being allowed to add other factors to the existing set. In addition, the tutors had the opportunity to revise these values during a *post-task walkthrough*.

The student screen was video-recorded during each session for the purpose of *replay* and *post-task walkthroughs* with the tutors. Immediately after each session, tutor and student were interviewed using a *semi-structured interview protocol*.

The data used for the analysis presented here consist of the dialogues and the (updated) situational factors rather than the more qualitative data from the interviews. However, as it is obvious, the qualitative data (see, Porayska-Pomsta et al., 2008) also shaped the following investigation.

6.6.3 Dialogue Analysis: Annotation and types of student actions

6.6.3.1 Introduction

In order to be able to employ machine learning to predict tutor diagnoses of the student affective states, it was necessary to have a handle on the utterances collected. Although investigating how the actual student utterances are related to their affect would be interesting, it was out of the scope of this thesis. Therefore, the analysis focuses on the communicative goals of the student utterances rather than the exact surface form. To accomplish that, the 26 tutorial interactions were annotated with this goal in mind.

The predominant pattern found in the dialogues collected is that of question-answer-feedback cycles where the tutor states a question, the student answers and the tutor follows up with appropriate feedback. This pattern has been described before by other researchers, e.g., (Dillon, 1990; Graesser, 1993, p.15). In particular, Graesser elaborates on the one-to-one tutorial interactions by providing a five-step dialogue frame where

1. the tutor is asking a question

2. the student is answering

3. the tutor provides feedback on answer

4. the tutor and the student are improving the quality of the answer

5. the tutor assesses student's understanding of answer

This five-step frame also applies to the tutorial interactions observed in this study. However, the definitions of the different steps require further refinement in order to describe the data. For example, the definition of step 3 needs to be expanded to take into account the fact that sometimes, at least in domains such as calculus, instead of full and final answers students may provide partial answers which are just steps towards the final solution. This may affect the nature of the feedback which, instead of being simply negative or positive, may have a more complex function, such as hinting or prompting. In some cases the tutor may even choose not to provide feedback at all until the student specifies all steps towards a desired answer.

Furthermore, due to the nature of the exercises in this domain, which involved refinement of the formulae (i.e. simplification), a tutor question usually leads to a series of student actions rather than to a single, final response. Such series of student actions typically form interaction blocks that are bounded by the initial tutor question and the student's final answer, and relate to the same exercise or problem. The dialogues contain many such blocks which, in line with similar observations, e.g., (Fox, 1993, p.21), are composed of a multitude of different types of student and tutor actions. For example, students often request confirmation of the correctness of their answers. Tutors asking gauging questions or expressing their assessment of student progress can result in interaction blocks with multiple levels of embedding. This relates to the rule of adjacency pairs (Wetherell et al., 2001) which refers to sequences of two (or more) utterances as being conditionally relevant.

In order to proceed and actually derive an annotation scheme the relevant literature was researched. A number of proposals exist that describe the structure of tutorial dialogues. Most of these look into the issue of student actions from the tutor perspective and concentrate more on understanding and classifying tutorial strategies and tactics or investigate the issue from the perspective of building computational models (e.g., Grosz and Sidner, 1986). A few schemes attempt to describe both tutor and student actions (Graesser et al., 1992; Graesser and Person, 1994; Stevenson, 1991), while others separate the characteristics of student actions (e.g., Shah et al., 2002).

The existing schemes are very useful for clarifying the categories of different actions in tutorial interactions. However, it is inherently difficult for an individual scheme to be suitable for capturing all types of dialogues that are collected in different educational contexts and subject domains. It is worth mentioning here that while this research was on going Dzikovska et al. (2006) present an annotation scheme based on the same dataset employed in this thesis. They also highlight that other models do not cover the full breadth of the utterances of this dataset. While their goals are slightly different and their classification is more coarse-grained, some similar classes exist. The annotation described here was primarily based on the classifications described in Graesser et al. (1992) and Shah et al. (2002), and took into account the work of Stevenson (1991). Additional types were added for the many cases that these schemes could not cover. The introduction of new types led to different higher level groupings and revised subtypes from the original schemes. These are presented in Section 6.6.3.3 after briefly describing the methodology behind the annotation.

6.6.3.2 Annotation Methodology

All 26 interactions, consisting of 340 actions with only mathematical formulae and 352 with text and mathematical formulae, were annotated. All tutor-student action pairs were collected and grouped according to coarse-grained categories such as questions, statements, repairs and simple answers.

The annotation was carried out through an iterative process of refinement and combination of the existing classification schemes by Graesser et al. and Shah et al. and by devising new classes where no existing category fitted.

The initial 9 broad categories (with up to three subtypes each) were used by two annotators to annotate a set of 6 representative dialogues from tutorial interactions, to check for consistency. Cohen's Kappa (Cohen, 1960) was calculated as an indication of the inter-annotator agreement. The first run yielded a Kappa of .741. It was evident that there were some misunderstandings in the description of different types and an apparent need for an additional category. After discussion between the annotators, an apparent need for an additional category. After discussion between the annotators, the conflicting annotations were resolved and the annotation scheme and descriptions revised. A further set of three dialogues was annotated by both annotators independently: the new Kappa value was .845. The full set of dialogues was annotated to this revised scheme. The final classification consists of 10 broad categories.

6.6.3.3 Types of student actions

The 10 broad categories of student actions include: (1) confirmation or verification of the answer (as in Graesser et al., Shah et al. and Dzikovska et al.); (2) query or request (which also appear in all the above classifications but with different subtypes); (3) challenge (in Graesser et al. and Shah et al. as *expectational answer*); (4) clarification (in Dzikovska et al. and similar to Shah's conversational *repair*); (5) statement about self (appears also in Dzikovska et al. and in Graesser et al. as *assertion*, but here a more general definition is taken that includes statement about affect and the students' knowledge); (6) planning (appears in Dzikovska et al. as *task progression*); (7) self-correction (similar to Dzikovska's *edits*) ; (8) help acceptance; (9) acknowledgement and (10) reflection (the latter three categories appear also in Dzikovska et al. as *routine dialogue*).

These are further described in table 6.3 together with an example from each one.

Table 6.3: Annotation classes for student action types, with descriptions and examples

Type	Code	Description	Example
Confirmation: answer	ConfAnswer	Explicit or implicit verification or judgement of an answer.	*Student:* $15(x^3-3x)*4(x^2-1)$, *something like that?*
Confirmation: planning	ConfPlanning	Verifying correctness of the chosen method, a rule or that the steps taken are in the right direction	*Student: ok im going to have to use the rule twice, right?*
Query/Request: procedural	ProcQuery	Questions to the tutor about the method or process that would allow the student to continue	*Student: so i should find its derivative?*
Query/Request: knowledge	KnowRequest	Student is requesting some missing (or forgotten) information (e.g., a rule, or formula).	*Student: what is* $1/\cos(x)$*?*
Query/Request: help	HelpRequest	An explicit request for help	*Student: could you remind me of the next step?*
Planning: tutor initiated	TutorPlanning	Responses that relate to tutorial planning (next actions, deciding what to do next): can be as *positive* in the sense of accepting tutor's suggestions or *negative* when the student is being evasive.	*Tutor: Do you want to try something harder again or do you want to practise doing a few more of these ? Student: i'll attempt something harder...*

Planning: student initiated	StdPlanning	A student initiated action that relates to tutorial planning (next actions, deciding what to do next)	*Tutor: .. you really know how to use the chain rule. Student: thanks, shall I simplify that one?*
Self-correction: unprompted	StdCorrection	A correction of an answer by the student herself	*Student: oops the cos should be* $\cos^2$
Self-correction: prompted	TutCorrection	A correction by the student but when the existence of the mistake was identified by the tutor.	*Tutor: can you spot your mistake Student: yes... it should be* $3 * x^2$
Statement about self: affect	AffectStmt	An utterance with the goal of expressing affect either explicitly or covertly. Further annotated as *negative* or *positive*.	*Student: i hate ones with fractions! here goes*
Statement about self: knowledge	KnowStmt	A statement about student's knowledge (even partial). Usually tutor driven, occurring at the beginning of the tutorial when tutor asks about student's knowledge. Includes student-initiated cases.	*Tutor: Why did you do that? Student: because i thought you multiplied the powers when they were bracketed*
Statement about self: stuckness or lack of knowledge	NoKnowStmt	Admitting lack of knowledge or being stuck.	*Student: i really dont know what i should be doing with this equation*

Challenge	StdChallenge	Student reflects some disagreement with the tutor or expresses strong belief on an answer.	*Student: why is it not* $cos(3x^2+3)$ *?*
Clarification	Clarification	Usually in response to tutor saying something that the student does not understand immediately: not necessarily due to lack of knowledge, may be due to language ambiguity.	*Student: What do you mean in the form of a question?*
Accept Help	HelpAccept	Student accepting tutor's proposal for providing help	*Tutor: Do you want a clue? Student: yes please*
Acknowledgement	Acknowledge	A statement acknowledging what the tutor said, manifesting agreement or understanding. Usually accompanied by an answer.	*Student: I get it* $2\cos x - \sin x$
Reflection	Reflection	Explicit reflection on previous moves by restating or revising knowledge just learnt, or by providing verbose answers indicating that the student is reflecting on what previously learnt.	*Student: Is that the chain rule used 3 times?*

6.6.4 Machine Learning Methodology

WEKA's J4.8 algorithm (described in Section 3.4.1) was used again for the tree induction. In order to train the decision tree, the algorithm is presented with instances automatically constructed and pre-processed from the raw data. Similar to the instances presented to the algorithm for the analysis of the data of the student perspective, the instances consist of a vector of features and an appropriate nominal class that will be described below. Fast Correlation Based Filtering is used again to remove attributes that do not seem to offer anything in the prediction task and for the evaluation stratified cross-validation is used. However, a few changes in the methodology followed during the study with students (see Section 6.2.3) are needed due to the nature of the data collected in this case.

- Due to the large size of the data (compared to the student-perspective data), the number of instances where the factors do not change is larger than the number of instances where the factors change.

  As before, two sets are presented to the training algorithm: one using only the instances where the factor values change (henceforth referred to as `Tut_Set_A`) and one with all the instances (`Tut_Set_AB`). Although the reasons a factor changes are of more importance, there is valuable information to be derived from the lack of change. Using the two sets of instances, both changes and lack of changes can be taken into account. Therefore, two different decision trees are built for each factor. These trees could be aggregated to increase the accuracy of the overall model. When performed automatically the technique is referred to as bootstrap aggregating or bagging (Breiman, 1996). At this stage this is done manually and only as an initial investigation to analyse the data and facilitates generation of fine-grained hypotheses for future research.

- In the study of student perspective, data were collected in advance as changes. Here the factor values were more fine-grained. While this provides richer data it complicates the training for the decision tree. For example, changes from `low` to `medium` would be treated differently from changes from `medium` to `high`. In this way, fewer instances exist of the same characteristics making it harder for the algorithm to generalise. Therefore rather than the exact factor values (low, medium, etc.) a class called `factor-change` takes values which encode its relative change. Accordingly, the class can take the following values:

1. `decrease, increase` which encodes the direction of the factor value change,
2. `extr_decrease, extr_increase` which encode changes in either direction between the extremes of the scale,
3. `init_medium, init_low, init_high` which encode the fact that the value is initialised from `irrelevant` to values in the corresponding region.
4. `no_change` is used in `Tut_Set_B` to depict the fact that a factor value was not changed .

6.6.5 Results

6.6.5.1 Confidence

There were 110 changes to the value of confidence (these comprise `Tut_Conf_Set_A`) in a set of 539 instances (`Tut_Conf_Set_AB`). As explained in the previous Section the machine learning algorithm is presented with both sets and two different trees can be generated. These are presented in Appendix D.3. A representation of the tree that results from `Tut_Conf_Set_A` when using attributes `action`, `difficulty` and `hist-vector` is shown graphically in Figure 6.3. In this case 61.818% of the cases are classified correctly and ($Kappa = .439$) which is interpreted as a 'fair' result (see Fleiss,1981). The tree generated based on `conf-setAB` classifies correctly 73.724%. It has a poor $Kappa = .316$ and, as expected, it is better at predicting non-changes (F-measure for `static` 0.853). However, it can be still used to provide support or to contradict the interpretations from the first tree.

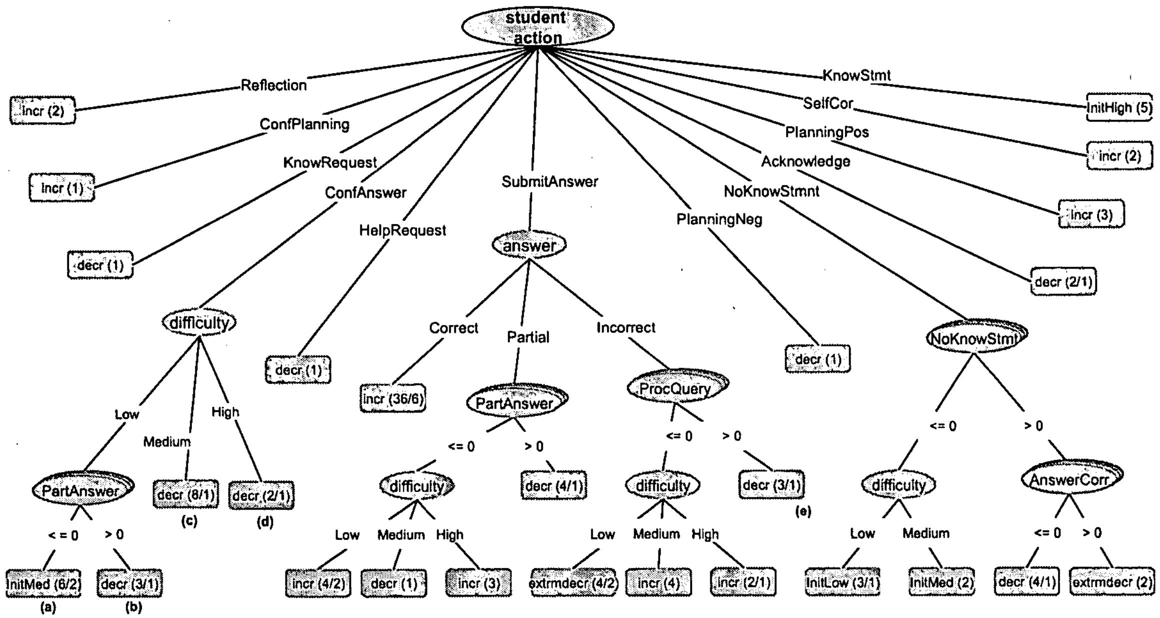


Figure 6.3: Decision tree for `confidence` trained on `Tut_Conf_Set_A`. The labelled nodes and leaves mark examples which are discussed in this section.

The two different trees offer clear indications of the underlying patterns and compliment or contradict each other. This way they provide useful results either for direct inclusion in a model (when the evidence is enough) or for generating hypotheses for future research.

For example, the path ending with leaves `A,B,C,D` in Figure 6.3 may be used to infer the following rules: when the student is confirming their answer (*ConfAnswer*), and the difficulty of the question is medium or high then tutors decrease the value of confidence (leaves `C` and `D`). If the difficulty of the question is low and previously the student was given no partially correct answer (*PartAnswer* ≤ 0) then tutors initialise the value of confidence to medium (leaf `A` *InitMedium*). If the student has previously given partially correct answers (*PartAnswer* > 0) then tutors decrease the value of confidence (leaf `B`).

In Section 6.6.4 we discussed that there is a need to model the cases where the factor remains the same. The rules generated from `Tut_Conf_Set_AB` can compliment the rules from `Tut_Conf_Set_A` and vice versa. For example, a rule in the tree generated from `Tut_Conf_Set_AB` suggests that there are a large number of cases where tutors do not change the values of confidence after a confirmation request during very difficult exercises. The rule suggests that this happens when the difficulty is High (there are 14 instances). This challenges the rule in Figure 6.3 ending with leaf `D`. An investigation in the raw data reveals that the 2 cases from the branch of the tree are the cases where the difficulty of material was set to `MediumHigh` by the tutor which could be interpreted as either Medium or High. Similarly, complementary rules to `Tut_Conf_Set_AB` can be added by inspecting `Tut_Conf_Set_A`. For example a rule in `Tut_Conf_Set_AB` indicates that as long as students do not request help, even after incorrect answers, their confidence remains unchanged. However, the branch ending at leaf `E` in Figure 6.3 suggests that what influences tutors' decisions to change the value of confidence is the presence or absence of previous procedural queries and the difficulty of the question. It is through cases like this, that particularly interesting and complex hypotheses can be generated to guide us in designing particular situations for future data collection.

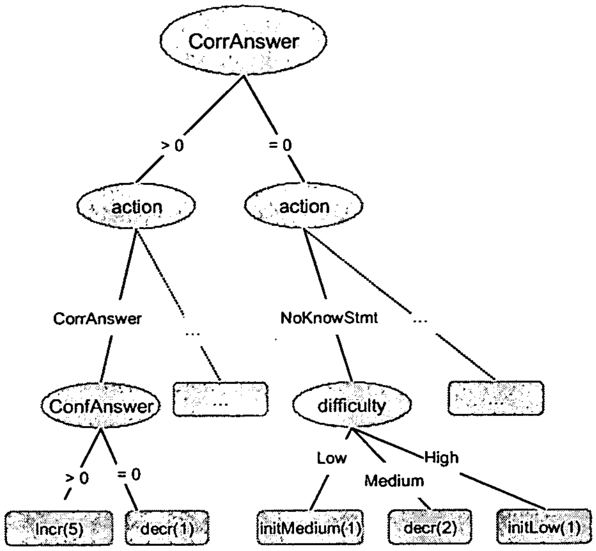


Figure 6.4: Excerpt from the decision tree for `effort` trained on `effort-set1`

6.6.5.2 Effort

There are 46 examples of changes (comprising `Tut_Effort_SetA`) and 326 examples of no changes (`Tut_Effort_Set_AB`) that were recorded for effort. The generated tree on `Tut_Effort_Set_A` classifies correctly only 30.4% of the instances and has a poor $Kappa = 0.3$. This is expected in such a small data set. The discriminating factor becomes one of the elements of the historical data, namely the number of correct answers prior to the change of the factor value (see a the fragment of the tree in Figure 6.4). Following the same approach as that in confidence, the decision tree based on the whole set of data is also generated (87.6712% instances classified correctly $Kappa = 0.3764$). Despite the poor results they can still be used to generate hypotheses for future data collection. For example, a rule suggests that when the difficulty of the question is in the Low or Medium region and the last and only action of the student is a declaration of their lack of knowledge, then tutors believe that students' effort is decreased or they initialised to indicate its importance. Another rule shows tendency to increase the value of student effort when the student submitted correct answers and in their previous correct responses they did not seek confirmation (see Figure 6.4). In another rule from the whole set of data, there are 3 cases where the diagnosed value for effort is decreasing immediately after a lack of knowledge statement in cases where there are no answer confirmations before that.

6.6.5.3 Interest

In the case of interest there are 32 instances where the factor changes (`int-set1`) out of the total 396 situations (`int-set2`). The tree trained on (`Tut_Int_Set_A`) correctly classifies 51.724% ($Kappa = .269$) and the tree trained on (`Tut_Int_Set_AB`) seems to overfit the data (95.0413% correctly classified instances and $Kappa = 0.6591$). These results may not be satisfactory again in terms of extracting an implementable model but they are informative for hypotheses generation.

Based on the generated trees, the actions that are associated with interest are queries about the process and the plan of the solution, self corrections and reflection. The conflicting cases in the two trees show that there are cases where the value of student interest increases. Unfortunately, the algorithm cannot differentiate between these and inspecting the processed data generates no apparent reasons. The raw data, in combination with the qualititative data from tutors' walkthroughs, indicate that, in these cases, student answers were particularly verbose. This is illustrated in the following examples.

Example 1

Tutor: A little bit harder this time. Differentiate $\sqrt{x^3 - 9 * x}$

Student: Right, i see

Student: I'll rewrite it first $(x^3 - 9 * x)^{1/2}$

Example 2

Tutor: Let's try a similar one: $\frac{1}{(\cos x^2 - 9)^{-1}}$?

Student: So rewrite as $(\cos x^2 - 9)$

Student: which equals $\frac{2 * \sin x^2 - 9}{\cos x^2 - 9^2}$

The walkthroughs suggest that multiple reflection actions by the student and their verbose and detailed answers made tutors increase the value of interest. Despite the sparsity of such instances, which currently prevents statistical validation of this hypothesis, the consistency of their occurrence is confirmed by comments by most tutors during the walkthroughs.

6.7 Discussion on Predictive Modelling From Tutor Perspective

The last two sections described two different methods of introducing tutors' perspective in the predictive modelling of student affect. The first was an attempt to elicit tutor's comments by showing them replays of students' interactions with the ILE. However, the pilot study conducted identified many difficulties that rendered the study infeasible. The most prominent of these was the choice of 'experts' in this context. As discussed in Section 6.5 the initial expectation that choosing tutors that are domain experts, use the ILE in classroom, and are supportive of its use did not ease the difficult task of observing an interaction in which they are not involved and the feel quite distant from. Watching these replays and inferring affective characteristics requires some skills that these two tutors do not necessarily have. It is hypothesised that (a) involving them in the interaction (through perhaps a wizard-of-oz type of experiment) or (b) selecting other experts (for example HCI experts familiar with tutoring aspects) would result in a more successful study.

In the second method a dataset consisting of tutors' affective diagnosis collected from empirical studies with tutors and students interacting with each other was analysed. The data lend themselves to interesting machine learning issues and analysis. First of all, because the data consisted of dialogues rather than a constrained set of actions (like in an ILE) they had to be annonated to facilitate the machine learning analysis.

The latter also raised important issues. First of all the different ways that they were collected required more pre-processing. Apart from the usual pre-processing (like missing values, noise etc.) that usually such data require, the fact that the values of the factors were fine-grained required regrouping them to values that depict the relative change of the factor. The methodology seems promising, however, due to the small set of the data, most of the rules here can only be treated as hypothesis for future studies.

Further investigation to improve the accuracy of the models revealed an interesting finding. The factor value prior to the change seems to have an impact on further changes. That is, tutors, depending on the region of the factor value, require different amount of evidence in order to change it. This is something that would not have been identified easily from a qualitative analysis or based on intuition. During the walkthroughs traces of this idea appear when different tutors mentioned that the reason they did not change easily the value of the factor (particularly to high value) despite evidence for that would otherwise be sufficient (e.g., the would indicate an increase from low to medium). This has important implications for the exact modelling approach and should be investigated further. A hypothesis that can be made is that, due to effects in the tutorial that such a diagnosis (e.g., high effort) would have, tutors are more reluctant to change the value. This is also characterised by the fact that the history of the interaction is more often used when the factor values are increasing. In a way even one action can make the tutors reduce the factor of a value but in order to increase it, the evidence has to build up over a series of actions.

A related issue is the initial value that one assumes students start the interaction with. One of the difficulties in the study with tutors watching students' interactions was the fact that they did not have enough information. Section 6.5 highlighted the fact that despite providing some information to the tutors about the student they still felt quite disengaged from the situation. It is interesting to observe that in the computer-mediated environment one of the first questions that tutors asked was aimed to probe into students' confidence.

Future data collection could concentrate particularly on these matters. An attempt was made to take into account the previous value of the factor under investigation in the predictions. Due to the small size of the data, however, including a variable which represents the prior value of the factor offers the highest information gain. Although the generated trees are more accurate (for example for effort the tree classifies correctly 63.043% of the instances and *Kappa* is 0.488) the sparsity of the data makes the machine learning algorithm produce degenerate rules, for example rules based on a single condition (e.g., when prior value is high then it is changed to medium) that do not offer much in terms of hypothesis generation or potential implementations because they are not sufficiently discriminatory.

As far as the analysis itself, it can also be improved by quantifying other aspects of the tutor-student interaction and including them in the analysis. A more detailed investigation of some of the misclassifications that occur, in conjunction with the relevant walkthroughs suggests that the value of factors also depends on aspects of the interaction that were not taken into account. For example, it is very clear from the walkthroughs that when students answer exceptionally quickly and straightforwardly (e.g., they use the equation editor competently) students increase the value of confidence. However, currently, the information about student's aptitude with using the equation editor are not captured in the data.

Chapter 7

General Discussion and Conclusions

7.1 Summary of results

The long-term objective of the research conducted in this thesis was to contribute towards the improvement of ILEs by understanding better and modelling students' behaviour, their emotions and motivation while they are working with them. Chapter 2 and 4 established that students use the features of ILEs in different ways than what they were designed for. This is particularly alarming when it comes to the feedback features of the system, which are designed to help students but are often used in ways that can be detrimental to their learning. It was also established that the issue is complex and is related to students' emotion and motivation, apart from their cognitive characteristics.

The research presented in this thesis achieved the following goals and results which can be separated in three parts. The first part was the design, redesign and implementation of components of WALLIS. Apart from the content and the feedback mechanism that was adapted to facilitate this research, the logging mechanism was a key requirement that enabled the recording of students realistic interactions. The second part relates to the better understanding of students' behaviour which was made possible thanks to the recordings of realistic data while students were interacting on their own time and location with a specific ILE (WALLIS).

- Factors that influence students' behaviour were investigated theoretically (in Section 2.4.3) and empirically, in relation to WALLIS (in Section 4.3). These facilitated the data interpretation and provided insight for the validity of the studies later on in the thesis.

- In Section 4.5 a better understanding of how students interact with WALLIS was gained. Several aspects of their overall behaviour were related to learning while particular attention was paid to their help-seeking behaviour. Its complex nature was appreciated and it was made clear that its relation to learning is not straightforward but relates to many other aspects that characterise it.

- Implications were drawn from the above which dictated a redesign of the system. This was presented in Chapter 5 and it involved, apart from HCI aspects of the system, the development of two Bayesian models. The first, predicts when students need help, and the second predicts if their interaction with the system is beneficial (in terms of learning). Apart from playing a role in adapting the feedback and interventions of the system, the models' predictions played a role at Chapter 6 of predictive features of affective diagnosis.

The second part investigated the use of machine learning, and specifically decision trees, for deriving predictive models of students' affective and motivational states.

- By recording students' interactions with WALLIS and eliciting self-reports about their own emotional and motivational factors, a decision tree was derived for each of the factors for which sufficient data were collected (namely *confidence* and *effort*) while hypotheses were derived for the rest of the factors that provide insight for future studies. These were presented in Section 6.2.4.

- During the above process two important issues were identified. First, that the cases during the interaction, for which there is no report, should also be taken into account. Second, that the history of the interaction plays a vital role in the predictions.

- Problems emerging because of the subjective nature of the data coming from self-reports were discussed in detail in Section 6.3. Although for some of the factors, students themselves are the best source of information, it seems that for the factors with negative connotation (e.g., reduced effort), students' reports tend to be biased.

- Given the previous finding, Section 6.4 established the need for a tutor perspective. An attempt was made to model tutors' diagnosis of the same factors by replaying students' interactions to them. The pilot study, presented in Section 6.5, was not entirely successful but due to the difficulty of the task for the tutors, the pilot study provided insights for future work.

- Another approach was attempted in Section 6.6 with data collected from a study designed to externalise tutors' diagnosis of students' characteristics while they were interacting with each other in a computer-mediated environment. Despite the small size of the data, the methodology employed for the analysis of the student perspective proved useful also in this case.

7.2 Contributions and recommendations

At first sight the tangible results mentioned previously may appear specific to the system employed in this thesis and the overall context. However, WALLIS resembles many state-of-the-art eLearning environments and the educational context within which it is integrated is very common, especially among higher education teaching and learning. Therefore, the findings contribute to the overall goal of improving ILEs.

More specifically, the in-depth analysis of students' interactions in Chapter 4 contributed further knowledge to a definitely weakly understood relationship between help-seeking and learning and a better understanding of which aspects of interactions are related to learning and how. The size of the data and the fact that some of the findings are repeated over the years support the evidence and suggest that they should be generalisable, at least in the context of higher education. Although the exact Bayesian models (i.e. their probability tables and their structure) derived in Chapter 5 may not be immediately applicable to other systems, the variables that seem to play a role in them, provide insight for similar work in other systems. More importantly, the methods behind the development of the models can be replicated. In addition, the redesign presented innovative ways to employ the predictions and letting them guide the feedback that the system provides, without risking a preventative approach that could impede students' learning (for example, denying help when this is needed). Similarly, the methods of data collection and analysis throughout Chapter 6, as well as the difficulties met and challenges faced, provide useful methodology for further studies. In particular, the classification of student actions at Section 6.6.3 contributes directly to the area of tutorial dialogue by extending schemes that did not cover adequately all student actions. As it includes types that the machine learning analysis linked to affective and motivational states of students, this classification can provide insights for the design of systems that take into account student language apart from their actions. Finally, the decision trees derived in Chapter 6 contribute to the field by offering directly implementable rules for predicting confidence and effort, as well as further insight into a quite complex situation, providing hypotheses that can stimulate future research.

The thesis overall, contributes directly and indirectly to several other areas in the field and particularly to aspects pertaining to methodology. The following outline the main methodological contributions.

- The importance of the principle of *ecological validity* has been highlighted several times. The principle requires methods, materials and settings of a study to approximate real-life situations. The background established that more research is required on the specific ways students interact with ILEs. As they become more and more integrated in educational situations, recording and analysing realistic data from students' interactions will surely prove a useful methodology for understanding them better and consequently improving ILEs. This thesis demonstrated that this is possible. During the research process, it was made clear that when conducting such empirical studies, the results can be influenced by many factors that a researcher should take into account.

- The thesis also demonstrated that data mining and machine learning analysis techniques can be particularly useful in augmenting the results of qualitative research. While systems are integrated in pedagogical situations, vast amounts of data are collected. Employing appropriate techniques can help detect patterns and improve the knowledge that researchers have about students. The results can inform the design of the system's components in a bias-free and objective way.

  By employing the methodology developed for collecting and analysing data in this thesis, significant amount of time can be reduced in future work. Analysing qualitative data is not only prone to bias but is also very time consuming. The methods described here can be used to facilitate the qualitative analysis and to triangulate the results from other sources (e.g., interviews) making them more accurate. It was also observed that the output of models, such as the model of beneficial interaction and the need for help, can be helpful in providing measures that humans are not directly necessarily good at and therefore have the potential to be used in further qualitative analysis.

Apart from these overarching contributions, the thesis contributes to the field by identifying several issues that should be taken into account in similar research or continuation of the current one. These span across three aspects (a) the data collection process (b) the techniques behind analysing the data and (c) the interpretation of these results. These are presented in detail below.

First in terms of the data collection process:

- In order to be able to analyse a situation by employing advanced quantitative techniques, attention has to be paid during the data collection. In this thesis, for example, recording student interactions in video form would not be enough. The low level recording (keyboard and mouse movements, button clicks, link selections, typing in input boxes etc.) enabled not only the replay of these interactions (which could have been achieved with video of course) but also (a) the in-depth analysis of the students' behaviour, (b) the development of predictive models based on their actions and (c) the time-stamping of actions and the linking to affective and motivational states through machine learning. In addition, such rich data have the potential to enable researchers to conduct evaluations of certain aspects the system with simulated students. Having recorded realistic interactions, interventions or adaptation features of a system can be evaluated without the dangers (and time involved in) recruiting real students.

- The issue described in the previous section about the cases of 'no report' seems specific but is something that will always apply to similar data collected from think-aloud protocols. Not reporting a factor value change does not mean that this factor did not change. An approach of how to tackle this was employed in the last study where tutors were asked explicitly to click 'no change' to signify that nothing changed in the situation.

- Another important point to carry over to further research is that particular attention needs to be paid in reporting the context of the data collection itself. In the case of this thesis, this included, apart from information of the ability of the students (background, academic achievement), goals that they were asked to achieve, feedback given by the system, and pre and post assessment results. Having details of such context, can facilitate comparisons and the identification of differences between data-sets (using for example meta-analysis techniques).

- By collecting datasets that are representative of the actual evidence that the ILE can collect during modelling, the data are in the right granularity to enable the machine learning analysis. This guarantees that the research results are transferable to the actual situation being modelled and that the models developed are easily implementable since they are less abstract and do not need much reification. Similarly, the bandwidth of the available information plays an important role. In the study from the tutors and students interacting with each other, by restricting the bandwidth of interaction to a chat interface without visual or audio channels, tutors have to adapt their teaching to this limited channel of communication. This was essential for identifying tutors' inferences that are directly relevant to the learning environment.

In terms of the data analysis techniques employed:

- In Chapter 4, it became apparent that when looking into correlations with learning, averaging the results of students across skills can provide misleading results. It is important to test learning or performance against data that correspond to the skill that is tested. It is not very clear if particular care is taken on this issue in the field. A methodology that looks particularly at this, is called Learning Factor Analysis and has been gaining ground in recent years (Koedinger and Junker, 1999; Freyberger et al., 2004; Cen et al., 2007).

- The clustering technique employed in Chapter 4 was not as useful as anticipated in identifying groups of students and did not yield stereotypes that could be employed to design appropriate interventions. However, it was quite useful in selecting a representative sample of students to interview. Therefore, the technique could be employed as a facilitator of qualitative research in other studies.

- When analysing data from dialogues, a common approach is to look into the surface form of the dialogues to investigate the issue one is interested in. In Section 6.6 a slightly different approach was demonstrated where the dialogues are first classified in terms of their communicative goal. Apart from the fact that this facilitated the machine learning analysis at the desired level for the current research, it also helped in providing generalisable results since they are applicable beyond the context of dialogue-oriented systems as they refer to actions which can happen through other means (e.g., menus or buttons) rather than the exact form of the study.

Finally, in terms of the interpretation of the results the following issues are important:

- Some of the aspects that need to be modelled in the area seem to be better modelled when different perspectives are taken into account. The next section discusses this in more detail.

- Results from educational research in classroom or one-to-one situations are not always transferable in the context of ILEs. They should cautiously be taken into account before directly applying them. Empirical research in the actual context where ILEs are integrated, has the potential to provide more valid and useful results.

- Even when empirical research is conducted in the field, it seems necessary to take into account all aspects of the context within which the research is situated when interpreting and reporting results, as it plays an important role. Context could explain the contradictory results of several pieces of research in the field. This is suggested by other researchers as well. For example Martinez-Miron et al. (2003) argue for the inclusion of context as an important variable when investigating learning goals orientation in ILEs. In this thesis, in Section 4.3.3, the influence of the lecturer on the way the system was introduced in the educational situation was obvious. It influenced the students' interactions and consequently, it would have affected the results of any analysis performed.

  Therefore, reporting explicitly the characteristics behind each study (as often done in educational research) could facilitate a future meta-analysis Glass and McGaw (1981) that could accumulate the results across these different studies.

7.3 Outstanding Issues and Future Work

This section presents some of the outstanding issues of the current research and how future work can improve the findings as well as the methodologies employed. In addition, two issues that appear throughout the thesis and are worth investigating further are presented.

First, in relation to how the exact results and findings could be improved:

- In Section 5.3 it was revealed that logistic regression could offer better accuracy than the Bayesian networks for the predictions (see also Appendix D). Since it has the potential to be implemented in an easier way and provide the probabilistic framework needed to cover the uncertainty of the prediction, it is something to consider for future implementations.

- Other techniques could also improve the accuracy of the models. For example, instead of just trying learning Bayesian networks or decisions trees on the directly observable features, linear and higher order combinations of them could prove a more appropriate approach. Work that can provide insights of how to do this is described in Baker (2005) where the performance of the model is optimised by using machine learning on the features themselves.

  Similarly, although the choice of decision trees seems particularly suitable at this stage, as they provide human inspectable results that can generate hypotheses, other techniques should also be tested once more data are collected. For example, Hidden Markov Models (HMM) have the potential to provide more accuracy and incorporate the history of the interaction in an easier way. Therefore they provide an appropriate framework for the situation. Examples of their use are starting to appear in the field in general (e.g., Beal et al., 2007; Soller and Lesgold, 2003) and that of affective computing in particular (Fernandez, 1997; Nwe et al., 2003; Jeffrey and Woolf, 2006).

- In Chapter 6 the small size and sparsity of the data in certain cases prevented the decision tree algorithm from extracting unequivocal rules in both student and tutor perspectives. To improve the accuracy of each model apart from developing algorithms that could deal better with the sparsity of the examples presented to them, it seems that a microanalysis of misclassified rules is needed. This would need to be done in triangulation with qualitative data.

- Finally, the models for predicting students' need for help and whether their interaction is beneficial could be designed to improve their accuracy *while* students use the system. The whole process described in Section 5.3 could be automated so as results from the performance of all students are available to the system permiting a continuous machine learning.

In terms of how the methodology and data collection can be improved further:

- Section 6.3 described in detail the difficulties with collecting students' reports on their own emotional and motivational states. Students seem to provide more accurate reports for the factors that have positive connotations. Therefore, a better setup that minimises the involvement of the researcher should reduce the bias considerably. In addition, a solution is needed to record explicitly that a factor value does not change, rather than assume it from the lack of students' reports, which are not necessarily the same. For example, during the study, the continuous replays could be stopped at particular points where a pattern seems to be emerging (e.g., based on the data collected so far) and, similarly to the study with the tutor's perspective, the student could click a button if they feel that their affective characteristics have not changed.

- The problem of the small size and sparsity of the data throughout the thesis has to be addressed. Collecting data in a more regular and perhaps automatic manner rather than the time-consuming walkthroughs that were performed could increase the size of the data. In addition, one could focus on particular events. These could have been identified in advance given that patterns are now known and hypotheses have been derived. A possible approach is to engineer particular situations that could probe the exact circumstances under which some of the reported changes occur.

- In terms of the evaluations performed, since they influence the judgements on the validity of the results, the 'blind' stratified cross-validation which was used could benefit from being performed in a way that takes into account the nature of the data and the 'cost' of misclassifications. A more formal process could validate rules against test data comprising of judgements from both expert tutors and students themselves. Another approach would be to measure the model's error in terms of being detrimental to students' learning by predicting, for example, that they are confident while they are not (false positives). Further work could look into adapting the existing learning algorithms to deal with such issues.

- The long-term goal behind the investigation of the tutor's perspective was initially to compare and complement the predictive modelling from students' perspective. In some sense, the two different perspectives can be reconciled to derive a more accurate one. While this is beyond the scope of this thesis, glimpses of how this could be achieved appeared within the analysis of the data from tutor-student interactions where it seems that by manually aggregating the different trees a more accurate model can be derived. The issue of automatically aggregating models has been investigated in detail in the field of data mining (Williams, 1990; Vannoorenberghe, 2004). In addition, the needs behind merging different perspectives resemble the emerging requirements behind reconciling models in the field of ontologies (e.g., Klein, 2001; Ehrig and Sure, 2004). A particularly similar example is the work presented in Aroyo et al. (2006) where a user's and an expert's conceptual model are compared. Developing formal ways to perform such an endeavour is necessary as this would reduce the bias introduced by researchers' intuition. Insights of how this could be done appear in Agarwal et al. (2005).

- Alongside the issue of reconciling perspectives, it was already discussed that both student and tutor perspectives have their advantages and limitations in terms of modelling the situation. Another perspective that could improve the modelling process is the one that could come from observations of students' interactions from peers. One of the tutors' problem when observing students' interactions was that they felt disengaged from the situation. It was also hypothesised that the work in de Vicente (2003) did not meet these difficulties because the role of the tutor was played by postgraduate students, who perhaps could empathise better with the student being watched. Therefore, it seems that employing externalisations of diagnosis based on peer students could alleviate both the problems that arise from self-reports and that of tutors.

Finally, it is worth noting three issues that appear consistently throughout the thesis:

- The first is related to students' *reflection on the feedback* that is provided to them in relation to their own actions. The issue appeared many times throughout the thesis and particular attention should be paid to it in the future. Reflection is not only related to learning but it emerged as evidence during both students' and tutors' walkthroughs in relation to affective and motivational characteristics. Future research could investigate in more detail ways of predicting it accurately and taking it into account for interventions and feedback provided by the system.

- The second is related to the *metacognitive skills* of the students. In terms of the help-seeking behaviour, Chapter 4 highlighted the complex way that it is related to learning. In addition, it was hypothesised that the differences in the two modes of the distribution of help-seeking frequency are also related to metacognitive, affective and motivational characteristics of the student. Although, at this stage, this was not explicitly investigated, it also appeared during the students' walkthroughs and tutors' externalisations of diagnostic rules. While future research could concentrate on improving the accuracy of the predictions for affective and motivational states, these predictions could be used to investigate the relation of affect to students' help-seeking behaviour. Similarly, the model of beneficial interaction and the prediction of need for help can be revisited including the affective predictions.

- The *situated nature of the research* and the fact that context was pervasive in many aspects of the issues investigated in this research was also highlighted several times. In particular, it was also identified that even the classroom's general goal orientation could play a particular role in how students interact with an ILE. The tutors during the pilot of students' replays required more information about the goal behind the students' interaction with the system. It is the nature of the phenomena observed that necessitates perhaps that even the systems when immersed in educational situations should also take into account the context when modelling students, providing feedback and adapting their content.

To conclude, it is worth revisiting some of the related research described in Section 2.5 where the results of this research find their particular relevance. The situation is analogous to a puzzle the pieces of which are the different research outcomes. The models developed, particularly when improved by further data collection, could be utilised in guiding adaptation and interventions of ILEs. In relation to the actual detection of students' emotions and motivation, as also mentioned in D'Mello et al. (2006b), detection accuracies could be increased by implementing hybrid models which combine results such as the decision trees of this research with other models derived based either on theories (e.g., Jaques and Viccari, 2007), on qualitative research (e.g., de Vicente, 2003), or even on more intrusive technologies (Picard and Scheirer, 2001; Fernandez et al., 1999; Leon et al., 2005). The outputs of such models can provide diagnostic functionalities to other frameworks (for example Conati, 2002; Morales et al., 2006) that have the potential to accumulate evidence. In relation to further adaptation and feedback provision, the accurate detection can feed in other models (e.g., situational model Porayska-Pomsta, 2003) and contribute to dialogue systems (e.g., Callaway et al., 2006) of ILEs which require evidence to adjust appropriately the feedback they provide. In Chapter 5, the system was redesigned using as minimalistic an approach as possible, based on the model of beneficial interaction. However, further cycles of redesign should investigate how the affective predictions can guide other appropriate interventions and motivational planning techniques, such as the ones studied by de Vicente (2003); del Soldato (1993); Rebolledo-Mendez (2007); Jaques and Viccari (2007) and others. The methodology and results of the aforementioned research together with the ones presented in this thesis are starting to complete the puzzle forming a holistic solution that can engage students and ILEs in a more authentic, beneficial interaction.

Bibliography

Abela, D. (2002). Improving a web-based ITS with ANA; an animated agent. Master's thesis, The Edinburgh University, School of Informatics;Artificial Intelligence.

ADL (2001). Sharable content object reference model. version 1.2: The SCORM overview. http://www.adlnet.org.

Agarwal, P., Huang, Y., and Dimitrova, V. (2005). Formal approach to reconciliation of individual ontologies for personalisation of geospatial semantic web. In Rodriguez, M., editor, *GeoS 2005, LNCS 3799*, pages 195–210, Springer-Verlag, Berlin.

Aist, G., Kort, B., Reilly, R., Mostow, J., and Picard, R. (2002). Experimentally augmenting an intelligent tutoring system with human-supplied capabilities: Adding human-provided emotional scaffolding to an automated reading tutor that listens. In *ITS 2002 Workshop on Empirical Methods for Tutorial Dialogue Systems*, San Sebastian, Spain.

Aleven, V. and Koedinger, K. (2000). Limitations of student control: Do students know when they need help? In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems*, volume 1839 of *LNCS*, pages 292–303.

Aleven, V. and Koedinger, K. (2001). Investigations into help seeking and learning with a cognitive tutor. In *Papers of the AIED-2001 Workshop on Help Provision and Help Seeking in Interactive Learning Environments*, pages 47–58.

Aleven, V., McLaren, B., Roll, I., and Koedinger, K. (2004). Toward tutoring help seeking: Applying cognitive modeling to meta-cognitive skills. In Lester, J. C., Vicario, R. M., and Paraguacu, F., editors, *Proceedings of 7th International Conference on Intelligent Tutoring Systems, ITS 2004*, pages 227–239, Berlin: Springer Verlag.

Aleven, V., Stahl, E., Schworm, S., Fischer, F., and Wallace, R. (2003). Help-seeking and help design in interactive learning environments. *Review of Educational Research*, 77(3):277–320.

Ames, C. and Archer, J. (1988). Achievement goals in the classroom: Students' learning strategies and motivation process. *Journal of Educational Psychology*, 80(3):260–267.

Ames, C. A. (1992). Classrooms: goals, structures, and student motivation. *Journal of Educational Psychology*, 84:267–271.

Anderson, J. (1993). *Rules of the mind*. Hillsdale, NJ: Erlbaum.

Anderson, J. R., Corbett, A. T., Koedinger, K. R., and Pelletier, R. (1995). Cognitive tutors: Lessons learned. *The Journal of the Learning Sciences*, 4(2):167–207.

Andres, E., Brna, P., Van Labeke, N., Mavrikis, M., Morales, R., Pain, H., and Porayska-Pomsta, K. (2005a). Student model specification. EU Project Deliverable No D10, LeActiveMath Consortium.

Andres, E., Brna, P., Van Labeke, N., Morales, R., Pain, H., and Porayska-Pomsta, K. (2005b). Diagnostic functionalities. EU Project Deliverable No D30, LeActiveMath Consortium.

Arbreton, A. (1998). Student goal orientation and help-seeking strategy use. In Karabenick, S., editor, *Strategic Help Seeking Implications for Learning and Teachin*, pages 95–116. Mahwah, NJ: Lawrence Erlbaum Associates.

Arnone, M. P. and Grabowski, B. L. (1992). Effects on children's achievement and curiosity in learner control over an interactive video lesson. *Educational Technology Research & Development*, 40(1):15–27.

Aroyo, A., Denaux, R., Dimitrova, V., and Pye, M. (2006). Interactive ontology-based user knowledge acquisition: A case study. In Sure, Y. and Dominguez, J., editors, *The Semantic Web: Research and Applications, 3rd European Semantic Web Conference, ESWC 2006*, pages 560–574.

Arroyo, I. and Woolf, B. (2005). Inferring learning and attitudes with a bayesian network of log files data. In Looi, C.-K., McCalla, G., Bredeweg, B., and Breuker, J., editors, *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology (Proceedings of AIED-2005 Conference, 18-22 July, 2005, Amsterdam)*, pages 88–95, Amsterdam, Holland. IOS Press.

Baker, R. (2005). *Designing Intelligent Tutors That Adapt to When Students Game the System*. PhD thesis, CMU Technical Report CMU-HCII-05-104, Carnegie-Mellon University.

Baker, R., Corbett, A., and Koedinger, K. (2004a). Detecting student misuse of intelligent tutoring systems. In Lester, J. C., Vicario, R. M., and Paraguacu, F., editors, *Proceedings of 7th International Conference on Intelligent Tutoring Systems, ITS 2004*, pages 531–540, Maceio, Brasil. Berlin: Springer Verlag.

Baker, R., Corbett, A., Koedinger, K., and Wagner, A. (2004b). Off-task behavior in the cognitive tutor classroom: When students "game the system". In *Proceedings of ACM CHI 2004 Conference on Human Factors in Computing Systems, 24-29 April 2004, Vienna, Austria*, pages 383–390. ACM Press, New York.

Baker, R., Roll, I., Corbett, A., and Koedinger, K. (2005). Do performance goals lead students to game the system? In *Looi et al. (2005)*, pages 57–64.

Balacheff, N. and Sutherland, R. (1993). Epistemological domain of validity of microworlds, the case of Logo and Cabri-geometre. In Lewis, R. and Mendelshon, P., editors, *Proceedings of the IFIP TC3/WG3.3 Working Conference on Lessons from Learning*, pages 137–150.

Bartholome, T., Stahl, E., and Bromme, R. (2004). Help-seeking in interactive learning environments: Effectiveness of help and learner-related factors in a dyadic setting. *International Conference of the Learning Sciences.*

Bartholome, T., Stahl, E., Pieschl, S., and Bromme, R. (2006). What matters in help-seeking? a study of help effectiveness and learner-related factors. *Computers in Human Behavior*, 22:113–129.

Baylor, A., Shen, E., and Warren, D. (2004). Supporting learners with math anxiety: The impact of pedagogical agent emotional and motivational support. In *Workshop on "Social and Emotional Intelligence in Learning Environments", held at the 7th International Conference on Intelligent Tutoring Systems, ITS 2004*, Maceio, Brazil.

Beal, C., Mitra, S., and Cohen, P. (2007). Modeling learning patterns of students with a tutoring system using hidden markov model. In *Luckin et al. (2007)*, pages 238–245.

Beck, J. (2004). Using response times to model student disengagement. In *ITS2004 Workshop on Social and Emotional Intelligence in Learning Environments.*

Beck, J. (2005). Engagement tracing: Using response times to model student disengagement. In Looi, C., McCalla, G., Bredeweg, B., and Breuker, J., editors, *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology*, pages 88–95. Amsterdam: IOS Press.

Bishop, C. (2006). *Pattern recognition and machine learning*. Information science and statistics. New York : Springer.

Bloom, B. (1976). *Human Characteristics and School Learning*. New York: McGraw-Hill.

Blum, A. and Langley, P. (1997). Selection of relevant features and examples in machine learning. *Artificial Intelligence*, 97:245–271.

Borkowski, J. G. and Thorpe, P. (1994). Self-regulation and motivation: A life-span perspective on underachievemen. In *Self-regulation of learning and performance*, pages 45–73. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Bouckaert, R. (2004). Bayesian networks in Weka. Technical report, Computer Science Department University of Waikato.

Bransford, J., Brown, A., and Cocking, R. (1999). *How People Learn: Brain, Mind, Experience, and School.* National Academies Press.

Bratman, M. (1990). What is intention? In Cohen, P., Morgan, J., and Pollack, M., editors, *Intentions in communication*, pages 15–31. MIT Press, Cambridge, MA.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2):123–140.

Brewer, M. (2000). Research design and issues of validity. In Reis, H. and Judd, C., editors, *Handbook of Research Methods in Social and Personality Psychology.* Cambridge University Press.

Brooks, C., Greer, J., Melis, E., and Ullrich, C. (2006). Combining ITS and elearning technologies: Opportunities and challenges. In M.Ikeda, Ashley, K., and Chan, T.-W., editors, *Proceedings of the 8th International Conference on Intelligent Tutoring Systems*, volume 4053 of *LNCS*, pages 278–287. Springer-Verlag.

Brown, P. and Levinson, S. C. (1987). *Politness: Some universals in language usage.* Number 4 in Studies in Interactional Sociolinguistics. Cambridge University Press.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *Journal of User Modeling and User Adapted Interaction*, 6(2-3):87–129.

Buntine, W. (1996). A guide to the literature on learning probabilistic networks from data. *IEEE Transactions on Knowledge and Data Engineering*, 8:195–210.

Burleson, W. and Picard, R. (2004). Affective agents: Sustaining motivation to learn through failure and a state of stuck. In *Social and Emotional Intelligence in Learning Environments Workshop in conjuction with the 7th International Conference on ITS.*

Callaway, C., Dzikovska, M., Matheson, C., Moore, J., and Zinn, Z. (2006). Using dialogue to learn math in the leactivemath project. In *ECAI Workshop on Language-enhanced Educational Technology*, Riva del Garda, Italy.

Cen, H., Koedinger, K., and Junker, B. (2007). Is over practice necessary? Improving learning efficiency with the cognitive tutor through educational. In *Luckin et al. (2007)*, pages 511–518.

Chen, H.-M. (2000). *An analytical approach to deriving usage patterns in a Web-based information system*. PhD thesis, School of Information Management and Systems, University of California.

Chi, M. T. H., Siler, S. A., Jeong, H., Yamauchi, T., and Hausmann, R. G. (2001). Learning from human tutoring. *Cognitive Science Society*, 25(4):471–533.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20:37–46.

Conati, C. (2002). Probabilistic assessment of user's emotions in educational games. *Journal of Applied Artificial Intelligence*, 16(7-8):555–575.

Conati, C., Gertner, A., VanLehn, K., and Druzdzel, M. (1997). On-line student modeling for coached problem solving using bayesian networks. In Jameson, A., Paris, C., and Tasso, C., editors, *User Modeling: Proceedings of the Sixth International conference*, New York: Spring Wien.

Conati, C. and Maclaren, H. (2004). Evaluating a probabilistic model of student affect. In Lester, J. C., Vicario, R. M., and Paraguacu, F., editors, *Proceedings of 7th International Conference on Intelligent Tutoring Systems, ITS 2004*, pages 55–66, Berlin: Springer Verlag.

Conati, C., Marsella, S., and Paiva, A. (2005). Affective interactions: the computer in the affective loop. In *Workshop at 2005 International Conference on Intelligent User Interfaces*. ACM Press.

Conklin, J. (1987). Hypertext: An introduction and survey. *IEEE Computer*, 20(9):17–41.

Conlon, T. and Pain, H. (1996). Persistent collaboration: a methodology for applied AIED. *International Journal of Artificial Intelligence in Education*, 7:219–252.

Coombs, N. (2006). Reflections on the state of accessibility in e-learning in the sciences. also available online http://rasem.nmsu.edu/Pdfs/symposium/coombs_rasem.pdf last accessed 06 Sep 2006.

Cooper, G. F. and Herskovits, E. (1992). A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–347.

Corbett, A. and Anderson, J. (1992). Student modeling and mastery learning in a computer-based programming tutor. In Frasson, C., Gauthier, G., and McCalla, G., editors, *Proceedings of the Second international conference on intelligent tutoring systems, ITS, 1992*. New York: Springer-Verlag.

Cozman, F. (2000). Generalizing variable elimination in bayesian networks. Technical report, Escola Politecnica, University of Sao Paulo.

Damasio, A. R. (1994). *Descartes' Error*. New York: Picador.

Damasio, A. R. (1999). *The feeling of what happens: body, emotion and the making of consiousness*. London: Vintage.

De Bra, P. and Calvi, L. (1998). AHA: a generic adaptive hypermedia system. In *In Proceedings of the 2nd Workshop on Adaptive Hypertext and Hypermedia*, pages 5–12, Pittsburgh.

de Vicente, A. (2003). *Towards Tutoring Systems that detect students' motivation: an investigation*. PhD thesis, The University of Edinburgh, ICCS, School of Informatics.

Dehn, D. and van Mulken, S. (2000). The impact of animated agents: a review of empirical research. *International Journal of Human-Computer Studies*, 52:1–22.

del Soldato, T. (1993). *Motivation in Tutoring Systemsn*. PhD thesis, School of Cognitive & Computer Sciences. University of Sussex. Brighton, UK.

del Soldato, T. and du Boulay, B. (1995). Implementation of motivational tactics in tutoring systems. *Journal of Artificial Intelligence in Education*, 4(6):337–378.

Dewey, J. (1913). *Interest and effort in education*, volume 7 of *Collected works of John Dewey. The middle works of John Dewey*. Carbondale, IL, Southern Illinois University Press,.

Dillon, A. and Gabbard, R. (1998). Hypermedia as an educational technology: A review of the quantitative research literature on learner comprehension, control, and style. *Review of Educational Research*, 68:322–349.

Dillon, J. T. (1990). *Practice of questioning*. Routledge: London.

D'Mello, S., Craig, S., Sullins, C., and Graesser, A. (2006a). Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16(1):3–28.

D'Mello, S. and Graesser, A. C. (2007). Mind and body: Dialogue and posture for affect detection in learning environments. In *Luckin et al. (2007)*, pages 161–168.

D'Mello, S. K., Craig, S., Sullins, J., and Graesser, A. C. (2006b). Predicting affective states expressed through an emote-aloud procedure from autotutor's mixed-initiative dialogue. *International Journal of Artificial Intelligence in Education*, 16:3–28.

Dochy, F. (1996). *Alternatives in Assessment of Achievements, Learning Processes and Prior Knowledge*, chapter Assessment of domain-specific and domain-transcending prior knowledge: Entry assessment and the use of profile analysis. London: Kluwer Academic Publishers.

Draper, S., Brown, M., Edgerton, E., Henderson, F., McAteer, E., Smith, E., and Watt, H. (1994). Observing and measuring the performance of educational technology, TLTP project.

du Boulay, B. and Luckin, R. (2001). Modelling human teaching tactics and strategies for tutoring systems. *International Journal of Artificial Intelligence in Education*, 12(3):235–256.

Duffy, T. M. and Jonassen, D. H., editors (1992). *Constructivism and the technology of instruction: A conversation*. Erlbaum, Hillsdale NJ.

Dweck, C. (1999). *Self-Theories: Their Role in Motivation, Personality, and Development*. Philadelphia, PA: Psychology Press.

Dzikovska, M., Callaway, C., Stone, M., and Moore, J. (2006). Understanding student input for tutorial dialogue in procedural domains. In *Brandial 2006: The 10th Workshop on the Semantics and Pragmatics of Dialogue*, Potsdam, Germany.

Eccles, J. and Wigfield, A. (1985). Teacher expectations and student motivation. In *Teacher Expectancies*, pages 185–226. Lawrence Erlbaum Associates.

Ehrig, M. and Sure, Y. (2004). Ontology mapping - an integrated approach. In *European Semantic Web Symposium (ESWS*, pages 76–691.

Eklund, J. (1995). Cognitive models for structuring hypermedia and implications for learning from the world wide web. In Debreceny, R. and Ellis, A., editors, *In Proceedings of Ausweb95: Innovation and Diversity*, pages pp. 111–117. Ballina, New South Wales: Norsearch Limited.

Ericsson, K. A. and Simon, H. A. (1993). *Protocol Analysis: Verbal Reports as Data*. MIT Press, Cambridge, MA.

Eugenio, B. D., Lu, X., Kershaw, T., Corrigan-Halpern, A., and Ohlsson, S. (2005). Positive and negative verbal feedback for intelligent tutoring systems. In *Proceedings of the 12th International Conference on Artificial Intelligence in Education*, Amsterdam, The Netherlands.

Felder, R. and Silverman, L. (1988). Learning and teaching styles in engineering education. *Engineering education*, 78:674–681.

Fernandez, R. (1997). Stochastic modeling of physiological signals with hidden markov models: A step toward frustration detection in human-computer interfaces. Master's thesis, MIT Thesis for the degree of M.S. in EECS.

Fernandez, R., Scheirer, J., and Picard, R. (1999). Expression glasses a wearable device for facial expression recognition. Technical report, MIT Media Lab.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions.* John Wiley & Sons.

Forbes-Riley, K. and Litman, D. (2004). Predicting emotion in spoken dialogue from multiple knowledge sources. In *HLT-NAACL 2004, Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts.

Forbus, K. and Feltovich, P., editors (2001). *Smart machines in education.* AAAI Press/The MIT Press, Cambridge, MA.

Forster, F., Hounsell, D., and Thompson, S. (1995). *Tutoring and Demonstrating : A Handbook.* Centre for Teaching, Learning and Assessment, The University of Edinburgh in association with the Universities' and Colleges' Staff Develoment Agency.

Fox, B. A. (1993). *The Human Tutorial Dialogue Project: Issues in the design of instructional systems.* Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Freyberger, J., Heffernan, N., and Ruiz, C. (2004). Using association rules to guide a search for best fitting transfer models of student learning workshop on analyzing student-tutor interaction logs to improve educational outcomes. In *Proceedings of 7th Annual Intelligent Tutoring Systems Conference.*

Gertner, A., Conati, C., and VanLehn, K. (1998). Procedural help in andes: Generating hints using a bayesian network student model. In *Artificial Intelligence AAAI-98*, pages 106–111. Cambridge, MA: The MIT Press.

Gillespie, R. (1991). *Manufacturing knowledge : a history of the Hawthorne experiments.* Cambridge : Cambridge University Press.

Glass, G. and McGaw, B. (1981). *Meta-analysis in social research.* Beverly Hills, CA: Sage.

Golden, G. (1990). Epistemology, constructivism, and discovery learning of mathematics. In R. Davis, C. M. and Noddings, N., editors, *Constructivist Views on the Teaching and Learning of Mathematics*, pages 31–47. Reston.

Goleman, D. (1996). *Emotional intelligence: Why it can Matter more than IQ.* Boomsbury, London.

Graesser, A. (1993). Dialogue patterns and feedback mechanisms during naturalistic tutoring. In *Proceedings of the Annual Conference of the Cognitive Science Society*.

Graesser, A. and Person, N. (1994). Question asking during tutoring. *American Educational Research Journal*, 31:104–137.

Graesser, A. C., Person, N., and Huber, J. D. (1992). Mechanisms that generate questions. In Lauer, T., Peacock, E., and Graesser, A. C., editors, *Questions and information systems*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Grosz, J. and Sidner, C. (1986). Attentions, intentions, and the structure of discourse. *Computational Linguistics*, 12(3):175–204.

Hall, M. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *International Conference on Machine Learning*, pages 359–366.

Heiner, C., Baker, R., and Kalina, Y., editors (2006). *Proceedings of the Educational Data Mining Workshop*.

Heiner, C., Heffernan, N., and Barnes, T., editors (2007). *Proceedings of Workshop on Educational Data Mining*.

Honey, P. and Mumford, A. (1986). *A Manual of Learning Styles,*. Maidenhead.

Howell, D. (1990). *Statistical Methods in Psychology*. Duxbury Press.

Hunn, C. (2003). Employing JESS for a web-based ITS. Master's thesis, The University of Edinburgh, School of Informatics.

Hunn, C. and Mavrikis, M. (2004). Improving knowledge representation tutoring and authoring in a component based ILE. *Poster appearing in Intelligent Tutoring Systems 2004*, 4(1):827–829.

Hunt, D. and Lawson, D. (1996). Trends in the mathematical competency of a level students on entry to university. *Teaching Mathematics and its Applications*, 15(4).

Issroff, K. (1996). *Investigating Computer-supported Collaborative Learning from an Affective Perspective*. PhD thesis, Institute of Educational Technology.

Jameson, A. (1996). Numerical uncertainty management in user and student modeling: An overview of systems and issues. *User Modeling and User-Adapted Interaction*, 5:193–251.

Jaques, P. and Viccari, R. (2007). A BDI approach to infer student's emotions. *Computers and education*, 49(2):360–384.

Jeffrey, J. and Woolf, B. (2006). A dynamic mixture model to detect student motivation and proficiency. In *Proceedings of the Twenty-first National Conference on Artificial Intelligence (AAAI-06)*, Boston, MA.

Joyes, G. (2006). Bridging cultures in designing for learning: An echina project case study. In *Proceedings of the Networked Learning Conference 2006.*

Kapoor, A. and Picard, R. (2005). Multimodal affect recognition in learning environments. In *Proceedings of the 13th annual ACM international conference on Multimedia (MM'05), 6-11 November, 2005, Singapore*, pages 677–682. ACM Press.

Karabenick, S. (1988). Help seeking as a strategic resource. In *Strategic Help Seeking Implications for Learning and Teaching*. Lawrence Erlbaum Associates.

Karabenick, S. (2003). Seeking help in large college classes: A personcentered approach. *Contemporary Educational Psychology*, 28:37–58.

Kardash, C. M. and Scholes, R. J. (1996). Effects of preexisting beliefs, epistemological beliefs, and need for cognition on interpretation of controversial issues. *Journal of Educational Psychology*, 88(2):260–271.

Keller, J. M. (1983). Motivational design of instruction. In Reigeluth, C. M., editor, *Instructional-design Theories and Models: an Overview of their Current Status*, pages 383–434. Laurence Erlbaum Associates Hillside New Jersey.

Klein, M. (2001). Combining and relating ontologies: an analysis of problems and solutions. In Gomez-Perez, A., Gruninger, M., Stuckenschmidt, H., and Uschold, M., editors, *Workshop on Ontologies and Information Sharing, IJCAI'01*, Seattle, USA.

Knowles, M. (1975). *Self-Directed Learning*. Chicago: Follet.

Knowles, M., Holton III, E., and Swanson, R. (1998). *The Adult Learner*. Butterworth:Heinemann, Woburn, MA.

Koedinger, K. R. (2001). Cognitive tutors as modeling tool and instructional model. In Forbus, K. and Feltovich, P., editors, *Smart machines in education: the coming revolution in educational technology*. AAAI Press/The MIT Press, Cambridge, MA.

Koedinger, K. R. and Junker, B. (1999). Learning factors analysis: Mining student-tutor interactions to optimize instruction. In *Social Science Data Infrastructure*, New York University.

Kohavi, R. and John, G. (1997). Wrappers for feature subset selection. *Artificial Intelligence*, 97:273–324.

Kolb, D. (1984). *Experiential Learning*. Prentice Hall.

Kort, B., Reilly, R., and Picard, R. (2001). An affective model of interplay between emotions and learning: Reengineering educational pedagogy; building a learning companion. In Okamoto, T., Hartley, R., Kinshuk, and Klus, J., editors, *International Conference on Advanced Learning Technologies*.

Krouse, J. and Krouse, H. (1981). Toward a multimodal theory of underachievement. *Education Psychologist*, 16:151–164.

Krutchen, P. (1999). *Rational Unified Process - an introduction*. Addison - Wesley, Reading, Massachusetts.

Larman, C. and Basili, V. (2004). Iterative and incremental development: A brief history. *Computer, IEEE CS*, page 48.

LeAM (2003). LeActiveMath Consortium: Language-enhanced, user adaptive, interactive elearning for mathematics. http://www.leactivemath.org: accessed 30th June, 2007.

Leon, E., Clarke, G., Sepulveda, F., and Callaghan, V. (2005). Real-time physiological emotion detection mechanisms: Effects of exercise and affect intensity. In *27th Conference of the IEEE Engineering in Medicine and Biology Society*, Shanghai, China. Engineering in Medicine and Biology 27th Annual Conference. Proceedings of the 2005 IEEE.

Lepper, M. R. and Chabay, R. W. (1988). Socializing the intelligent tutor: Bringing empathy to computer tutors. In *Learning issues for intelligent tutoring systems*, pages 242–257. Springer-Verlag, New York.

Lepper, M. R. and Woolverton, M. (2001). The wisdom of practice: Lessons learned from the study of highly effective tutors. In Aronson, J., editor, *Improving academic achievement: Contributions of social psychology*, pages 133–156. Academic Press., Orlando, FL.

Lepper, M. R., Woolverton, M., Mumme, D. L., and Gurtner, J. (1993). Motivational techniques of expert human tutors: Lessons for the design of computer-based tutors. In Lajoie, S. and Derry, S., editors, *Computers as Cognitive Tools*, pages 75–107. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

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 *CHI '97: Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 359–366, New York, NY, USA. ACM Press.

Lewis, M. and Anderson, J. R. (1985). Discrimination of operator schemata in problem solving: Procedural learning from examples. *Cognitive Psychology*, 17:26–65.

Lin, J., Keogh, E., Lonardi, S., and Chiu, B. (2003). A symbolic representation of time series, with implications for streaming algorithms. In *Proceedings of the 8th ACM SIGMOD workshop on Research issues in data mining and knowledge discovery*, pages 2–11.

Litman, D.J.; Forbes-Riley, K. (2006). Recognizing student emotions and attitudes on the basis of utterances in spoken tutoring dialogues with both human and computer tutors. *Speech communication*, 28(5):559–590.

Lohr, S. (1999). *Sampling: Design and Analysis.* Duxbury Press.

Looi, C., McCalla, G., Bredeweg, B., and Breuker, J., editors (2005). *Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology (Proceedings of the 12th International Conference on Artificial Intelligence in Education, AIED2005)*, volume 125 of *Frontiers in Artificial Intelligence and Applications*, Amsterdam, Nederlands. IOS Press.

LTSN (2000). Measuring the maths problem. Technical report, The Learning and Teaching Support Network (Maths, Stats & OR), The Institute of Mathematics and its Applications, The London Mathematical Society, The Engineering Council.

Luckin, R., Koedinger, K., and Greer, J., editors (2007). *Artificial Intelligence in Education: Building Technology Rich Learning Contexts that Work (Proceedings of the 13th International Conference on Artificial Intelligence in Education, AIED2007)*, volume 158 of *Frontiers in Artificial Intelligence and Applications*, Los Angeles, California. IOS Press.

Luckin, R., Underwood, J., du Boulay, B., Holmberg, J.and Kerawalla, L., O'Connor, J., Smith, H., and Tunley, H. (2006). Designing educational systems fit for use: A case study in the application of human centred design for AIED. *International Journal of Artificial Inteligence in Education*, 16(353–380).

Maesschalck, D., Jouan-Rimbaud, D., and Massart, D. (2000). The mahalanobis distance. *Chemometrics and Intelligent Laboratory Systems*, 50:1–18.

Mahalanobis, P. (1936). On the generalised distance in statistics. In *Proceedings of the National Institute of Science of India*, volume 12, pages 49–55.

Malone, T. W. (1981). Towards a theory of intrinsically motivating instruction. *Cognitive Science*, 4:333–369.

Martin, J. and Vanlehn, K. (1995). Student assessment using bayesian nets. *International Journal of Human-Computer Studies*.

Martinez-Miron, E., Harris, A., du Boulay, B., Luckin, R., and Yuill, N. (2003). The role of learning goals in the design of ILEs: Some issues to consider. In *Artificial Intelligence in Education Conference*, Amsterdam, Netherlands.

Masthoff, J. and Gant, A. (2006). In pursuit of satisfaction and the prevention of embarrassment: Affective state in group recommender systems. *User Modeling and User-Adapted Interaction*, 16(3-4):281–319.

Mavrikis, M. (2001). Towards more intelligent and educational dynamic geometry environments. Master's thesis, Division of Informatics.

Mavrikis, M. (2004). Improving the effectiveness of interactive open learning environments. In *3rd Hellenic Conference on Artificial Intelligence*, pages 260–269.

Mavrikis, M. (2005). Logging, replaying and analysing students' interactions in a web-based ILE to improve student modelling. In *Looi et al. (2005)*, page 967.

Mavrikis, M. and Maciocia, A. (2002). Developing WALLIS; a web-based system to enhance mathematics teaching. In *Proceedings of the Second International Conference on the Teaching of Mathematics*, Crete, Greece.

Mavrikis, M. and Maciocia, A. (2003a). Incorporating assessment in a web-based ILE. *LTSN maths-CAA series*. Available online at http://ltsn.mathstore.ac.uk/articles/maths-caa-series/index.shtml.

Mavrikis, M. and Maciocia, A. (2003b). WaLLiS: a web-based ILE for science and engineering students studying mathematics. *Supplemental Proceedings of the 11th International Conference on Artificial Intelligence in Education, AIED2003, Vol. VIII, Workshop on Advanced Technologies for Mathematics Education*, pages 505–513.

Mavrikis, M. and Maciocia, A. (2006). WaLLiS: a Web-based ILE for Science and Engineering Students Studying Mathematics. In *Electronic Proceedings of WebALT 2006*.

Mavrikis, M., Maciocia, A., and Lee, J. (2007). Towards predictive modelling of student affect from web-based interactions. In *Luckin et al. (2007)*, pages 169–176.

Mayo, M. and Mitrovic, A. (2001). Optimising ITS behaviour with bayesian networks and decision theory. *International Journal of Artificial Intelligence in Education (IJAIED)*, 12(2):124–153.

McKendree, J., Stenning, K., Mayes, T., Lee, J., and Cox, R. (1998). Why observing a dialogue may benefit learning: The vicarious learner. *Journal of Computer Assisted Learning*, 14(2):110–119.

Melis, E., Andres, E., Franke, A., Frischauf, A., Goguadse, G., Libbrecht, P., Pollet, M., and Ullrich, C. (2001). ActiveMath: A web-based Learning Environment. *International Journal of Artificial Intelligence in Education (IJAIED)*, 12:385–407.

Melis, E. and Ullrich, C. (2003). Local and global feedback. In Hoppe, U., Verdejo, F., and Kay, J., editors, *Proceedings of the 11th International Conference on Artificial Intelligence in Education, AIED 2003*, pages 476–478. IOS Press.

Merrill, D., Reiser, B., Merrill, S., and Landes, S. (1995). Tutoring: Guided learning by doing. *Cognition and Instruction*, 13(3):315–372.

Messom, C., Sarrafzadeh, A., Johnson, M., and Chao, F. (2005). Affective state estimation from facial images using neural networks and fuzzy logic. In Wang, D., editor, *Neural Networks Applications in Information Technology and Web Engineering*.

Morales, R., van Labeke, N., and Brna, P. (2006). Approximate modelling of the multidimensional learner. In *Proceedings of the 8th International Conference on Intelligent Tutoring Systems. ITS 2006*, volume 4053 of *LNCS*, pages 555–564. Springer.

Mostow, J., Beck, J., Winter, S., Wang, S., and Tobin, B. (2002). Predicting oral reading miscues. In *International Conference on Spoken Language Processing*, pages 1221–1224, Denver, USA.

Moundridou, M. and Virvou, M. (2002). Evaluating the personal effect of an interface agent in a tutoring system. *Journal of Computer Assisted Learning*, 18(3):253–261.

Murray, R. and VanLehn, K. (2005). Effects of dissuading unnecessary help requests while providing proactive help. In *Proceedings of the 12th International Conference on Artificial Intelligence In Education*, pages 887–889., Amsterdam, The Netherlands. AIED 2005.

Murray, R. C. and VanLehn, K. (2000). DT tutor: A decision-theoretic, dynamic approach for optimal selection of tutorial actions. In *Proceedings of the 5th International Conference on Intelligent Tutoring Systems. ITS 2000*. Berlin: Springer-Verlag.

Nelson-Le Gall, S. (1981). Help seeking : An understudied problem-solving skill in children. *Developmental Review*, 1:224–246.

Nelson-Le Gall, S. (1987). Necessary and unnecessary help seeking in children. *Journal of Genetic Psychology*, 148:53–62.

Nelson-LeGall, S. (1985). Help-seeking behavior in learning. *Review of Research on Education*, 12:55–90.

Newman, R. (1998). Students' help seeking during problem solving: Influences of personal and contextual achievement goals. *Journal of Educational Psychology*, 90:644–658.

Newman, R. S. (1994). Adaptive help seeking: a strategy of self-regulated learning. In Schunk, D. and Zimmerman, B., editors, *Self-regulation of learning and performance: Issues and educational applications*, pages 283–301. Hillsdale, NJ: Erlbaum.

Nicholls, J., Cobb, P., Wood, T., Yackel, E., and Patashnick, M. (1990). Assessing students' theories of success in mathematics: Individual and classroom differences. *Journal for Research in Mathematics Education*, 21(2):109–122.

Nisbett, R. and Wilson, T. (1977). Telling more than we can know: Verbal reports on mental processes. *Psychological Review*, 84:231–259.

Nwe, T., Foo, S., and De Silva, L. (2003). Speech emotion recognition using hidden markov models. *Speech Communication*, 41(4):603–623.

O'Bryen, P. (1996). Using questionnaires to assess motivation in second language classroom. Technical report, University of Hawaii Working Papers in ESL.

Op't Eynde, P., De Corte, E., and Verschaffel, L. (2001). What to learn from what we feel? The role of students' emotions in the mathematics classroom. In Volet, S. and Jarvela, S., editors, *Motivation in Learning Contexts. Theoretical Advances and Methodological Implications*. Oxford, UK.

Ortony, A., Clore, G., and Collines, A. (1988). *The Cognitive Structure of Emotions*. Cambridge University Press.

Peters, T. (1998). Remotely familiar: Using computerized monitoring to study remote use. *Library Trends*, pages 7–20.

Philips, R. (2005). Challenging the primacy of lectures: The dissonance between theory and practice in university teaching. *Journal of University Teaching & Learning Practice*, 2:1–12.

Picard, R. (1997). *Affective Computing*. MIT Press.

Picard, R. and Scheirer, J. (2001). The galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings of the 9th International Conference on HCI*, pages 1538–1542.

Pintrich, P. (2002). Future challenges and directions for theory and research on personal epistemology. In Hofer, B. and Pintrich, P., editors, *Personal epistemology: The psychology of beliefs about knowledge and knowing*, pages 389–414. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Porayska-Pomsta, K. (2003). *Influence of Situational Context on Language Production: Modelling Teachers' Corrective Responses.* PhD thesis, The University of Edinburgh.

Porayska-Pomsta, K., Mavrikis, M., and Pain, H. (2008). Diagnosing and acting on student affect: the tutor's perspective. *User Modeling and User-Adapted Interaction: The Journal of Personalization Research*, 18(1-2):125–173.

Porayska-Pomsta, K. and Pain, H. (2004). Exploring methodologies for building socially and emotionally intelligent learning environments. In *Proceedings of the Workshop on Social and Emotional Intelligence in Learning Environments (SEILE), ITS 2004*, Maceio, Brazil.

Puustinen, M. (1998). Help-seeking behavior in a problem-solving situation: Development of self-regulation. *European Journal of Psychology of Education*, 13:271–282.

Quinlan, J. (1993). *C4.5: Programs for Machine Learning.* Morgan Kaufmann, San Francisco, CA.

Quinlan, J. (1996). Improved use of continuous attributes in c4.5. *Journal of Artificial Intelligence Research*, 4:77–90.

Rebolledo, G., du Boulay, B., and Luckin, R. (2006). Motivating the learner: an empirical evaluation. In Ikeda, M., Ashley, K. D., and Chan, T.-W., editors, *Proceedings of Eighth International Conference on Intelligent Tutoring Systems, ITS 2006, June 26-30, 2006, Jhongli, Taiwan*, LNCS 4053, pages 545–554. Springer, Berlin.

Rebolledo-Mendez, G. (2007). *Modelling the learners motivational state in a Vygotskyan-inspired ITS.* PhD thesis, School of Cognitive & Computer Sciences. University of Sussex. Brighton, UK.

Renkl, A. (2002). Learning from worked-out examples: Instructional explanations supplement self- explanations. *Learning and Instruction*, 12:149–176.

Reye, J. (1998). Two-phase updating of student models based on dynamic belief networks. In Goettle, B., Halff, H., Redfield, C., and Shute, V., editors, *International Conference on Intelligent Tutoring Systems*, pages 274–283. Springer-Verlag.

Rollett, B. (1987). Effort avoidance and learning. In De Corte, E., Lodewijks, H., Parmentier, R., and Span, P., editors, *Learning and Instruction*, volume 1, pages 147–157. Oxford University Press.

Rozin, P. and Cohen, A. (2003). High frequency of facial expressions corresponding to confusion, concentration, and worry in an analysis of maturally occurring facial expressions of americans. *Emotion*, 3:68–75.

Russell, S. and Norvig, P. (1995). *Artificial Intelligence: A Modern Approach.* Los Altos CA, Morgan-Kaufman.

Scardamalia, M. (1992). The role of adaptation and understanding in knowledge-building communities. In *NATO Advanced Science Institute Conference*, Kolymbari Crete, Greece.

Schank, R. and Neaman, A. (2001). Motivation and failure in educational simulation design. In Forbus, K. and Feltovich, P., editors, *Smart machines in education: the coming revolution in educational technology.* AAAI Press/The MIT Press, Cambridge, MA.

Schiefele, U. (1991). Interest, learning, and motivation. *Educational Psychologist*, 26(3):229–323.

Schoenfeld, A. (1987). *Cognitive science and mathematics education*, chapter What's all the fuss about metacognition?, pages 189–215. Lawrence Erlbaum Associates, Hillsdale, NJ.

Schoenfeld, A. H. (1983). Beyond the purely cognitive: belief systems, social cognitions, and metacognitions as driving forces in intellectual performance. *Cognitive Science*, 7:329–363.

Schommer, M. (1993). Epistemological development and academic performance among secondary students. *Journal of Educational Psychology*, 85(3):406–411.

Schroder, M. and Cowie, R. (2006). ssues in emotion-oriented computing - toward a shared understanding. In *Workshop on Emotion and Computing - Current Research and Future Impact at KI 2006.*, Bremen, Germany.

Schworm, S. and Renkl, A. (2002). Learning by solved example problems: Instructional explanations reduce self-explanation activity. In Gray, W. D. and Schun, C. D., editors, *Proceedings of The 24th Annual Conference of the Cognitive Science Society*, pages 816–821, Mahwah, NJ.

Shafer, G. (1976). *A Mathematical Theory of Evidence.* Princeton University Press.

Shah, F., Evens, M., Michael, J., and Rovick, A. (2002). Classifying student initiatives and tutor responses in human keyboard-to-keyboard tutoring sessions. *Discourse Processes*, 33(1):23–52.

Sharples, M., Jeffery, N., du Boulay, B., Teather, D., and Teather, B. (1999). Socio-cognitive engineering. In *European conference on Human Centred Processes.*

Slavin, R. (1987). Ability grouping and student achievement in elementary schools: a best evidence synthesis. *Review of Educational Research*, 57(3):293–336.

Snow, R., Corno, L., and Jackson, D. (1996). Individual differences in affective and conative function. In Berliner, D. and Calfee, R., editors, *Handbook of educational psychology*. New York: Simon & Schuste.

Soller, A. and Lesgold, A. (2003). A computational approach to analyzing online knowledge sharing interaction. In Hoppe, U., Verdejo, F., and Kay, J., editors, *International Conference on Artificial Intelligence in Education*, pages 253–260. Amsterdam: IOS Press.

Solomon, B. A. and Felder, R. M. (1998). Solomon and felder's classification sequential vs global. Available online at http://www.ncsu.edu/felder-public/ILSdir/styles.htm.

Spensley, F., Elsom-Cook, M., Byerley, P., Brooks, P., Federici, M., and Scaroni, C. (1990). Using multiple teaching strategies in an its. In Frasson, C. and Gauthier, G., editors, *Intelligent Tutoring Systems: At the Crossroads of Artificial Intelligence and Education*, pages 188–226. Ablex Publishing Corporation, New Jersey.

Stevens, R., Soller, A., Cooper, M., and Sprang, M. (2004). Modeling the development of problem solving skills in chemistry with a web-based tutor. In *Proceedings of the 7th International Conference on ITS, Maceis, Brasil.*

Stevenson, J. C. (1991). Cognitive structures for the teaching of adaptability in vocational education. In Evans, G. T., editor, *Learning and teaching cognitive skills*, Monograph Series II, pages 144–163. Australian Council for Educational Research.

Stone, A., Bachrach, C., Jobe, J., Kurtzman, H., and S., H. (1992). *The Science of Self-report; Implications for Research and Practice*. Lawrence Erlbaum Associates.

Thuring, M., Hannemann, J., and Haake, J. (1995). Hypermedia and cognition: Designing for comprehension. *Communications of the ACM*, 38(8):57–66.

Tobias, S. (1994). Interest, prior knowledge and learning. *Review of Educational Research*, 64(1):37–54.

Ullrich, C., Wallach, D., and Melis, E. (2003). What is poor man's eye tracking good for. In O'Neill, E., Palanque, P., and Johnson, P., editors, *17th Annual Human- Computer Interaction Conference 2003*, volume 2, Swindon. British Computer Society.

VanLehn, K. (1988). *Foundations of Intelligent Tutoring Systems*, chapter Student modeling, pages 55–78. Hillsdale, NJ: Lawrence Erlbaum Associate.

VanLehn, K. and Martin, J. (1998). Evaluation of an assessment system based on bayesian student modeling. *International Journal of Artificial Intelligence in Education*, v. 8(2):179–221.

Vannoorenberghe, P. (2004). On aggregating belief decision trees. *Information Fusion*, 5(3):179–188.

Verma, T. and Pearl, J. (1992). An algorithm for deciding if a set of observed independencies has a causal explanation. In *Proceedings of the Eighth Conference on Uncertainty in Artificial Intelligence*, pages 323–330.

Weiner, B. (1990). History of motivational research in education. *Journal of Educational Psychology*, 82(4):616–622.

Weiner, B. (1992). *Human motivation: Metaphors, theories, and research.* Newbury Park, CA: Sage Publications.

Wetherell, M., Taylor, S., and Yates, S., editors (2001). *Discourse as data; A Guide for Analysis.* Sage and The Open University, London and Milton Keynes.

Williams, G. (1990). *Inducing and Combining Decision Structures for Expert Systems.* PhD thesis, The Australian National University.

Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, San Francisco, 2nd edition edition.

Wong, D., Clarke, S., Lodge, N., and Shephard, K. (2007). Demand-led e-learning and the elusive total solution. *British Journal of Educational Technology*, 38:116–132.

Wood, D. (2001). Scaffolding, contingent tutoring and computer-supported learning. *International Journal of Artificial Intelligence in Education*, 12:280–292.

Wood, H. and Wood, D. (1999). Help seeking, learning and contigent tutoring. *Computers & Education*, 33:153–169.

Woolf, B., Beck, J., Eliot, C., and Stern, M. (2001). Growth and maturity of ITS. In Forbus, K. and Feltovich, P., editors, *Smart machines in education: the coming revolution in educational technology.* AAAI Press/The MIT Press, Cambridge, MA.

Yu, L. and Huan, L. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *International Conference on Machine Learning*, Washington DC.

Zimmerman, B. (1994). Dimensions of academic self-regulation: A conceptual framework for education. In Schunk, D. and Zimmerman, B., editors, *Self-regulation of learning and performance: Issues and educational applications*, pages 3–21. Hillsdale, New Jersey: Lawrence Erlbaum Associates.

Zimmerman, B. and Pons, M. (1988). Construct validation of a strategy model of student self-regulated learning. *Journal of Educational Psychology*, 80:284–290.

Zukerman, I. and Albrecht, D. (2001). Predictive statistical models for user modeling. *User Modeling and User-Adapted Interaction*, pages 5–18.

Appendix A

Chronology of events

Figure A.1 below illustrates the chronology of the different activities undertaken as part of this thesis.

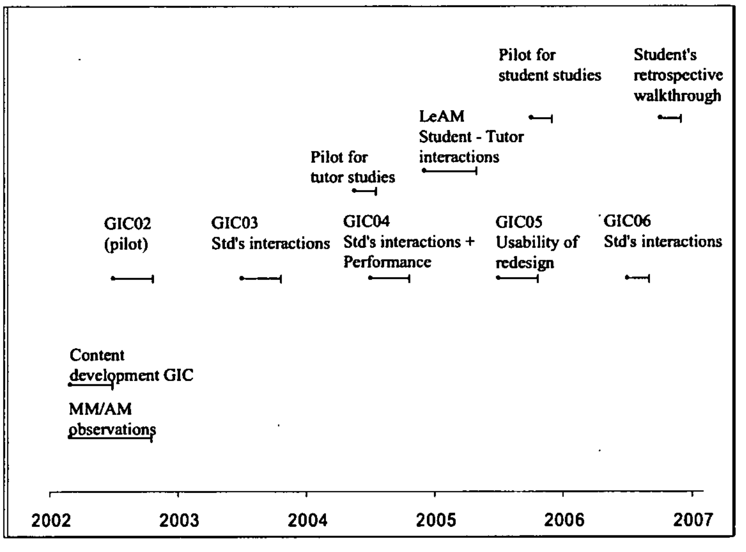


Figure A.1: Chronology of the different activities undertaken as part of the thesis.

Appendix B

Content in WALLIS

The following pages present the content written and the activities developed for this thesis as part of WALLIS for the *Conic Sections* part of the the *Geometry Iteration and Convergence* course of the School of Mathematics.

Conic sections page:***CS***

Apart from lines and planes in space there are other curves that arise from quadratic equations. These are called ***conic sections***.

The goal of these pages is to teach you:

- how conic sections arise geometrically and algebraically
- how these definitions relate to their graphs
- about the different types (ellipses, hyperbolae and parabolae).
- a more geometrical description of these curves; using special points and lines associated to the curves (the foci and directrices).
- the polar form of the ellipse
- and what the dual conic is

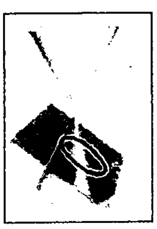
Geometric definition page:***GD***

Consider the set:

$$\{(x,y,z) \in \mathbf{R}^3 | \, x^2+y^2=z^2\}$$

This is a (double) cone centred on the origin. If we intersect this with a plane then we obtain a **conic section**.

For example:



- When we intersect with the plane z=1 we obtain the equation x^2+y^2=1 which is a circle (the plane in this case is horizontal).
- If the plane cuts only one of the cones and the resulting curve is **closed** (i.e. it does not go off to infinity) then the curve is an **ellipse** (see figure).
- If the plane cuts one cone but the curve is unbounded then we get a **parabola**.
- If the plane cuts both cones (but does not contain the origin) then we get two unbounded curves which are the two parts of a **hyperbola**.

You can see all this at the geometric definition activity

Algebraic definition

page:*AD*

Consider the following locus:

$$\{(x,y) \in \mathbf{R}^2 \mid \alpha x^2 + 2\beta xy + \gamma y^2 + \delta x + \varepsilon y + f = 0\}$$

where one of the α, β, γ must be non-zero.

This is the most general quadratic you can write down in the plane. It is a fact that all conic sections are solutions of such an equation and can be written in a simpler form called **standard form** that looks like the following :

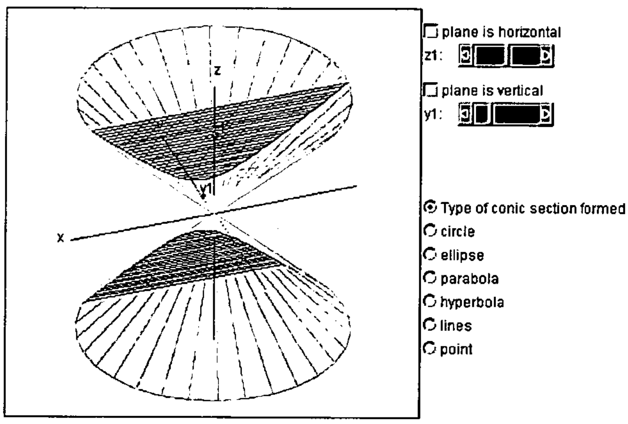
$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} = 1$$

This is a very useful format as it can be used to recognise and classify a conic

Geometric definition - activity

page:*GD-A*

This applet shows a double cone and its intersection with a plane. By *moving the sliders* you can change the angle with which the plane intersects the cone. Notice that with the two check boxes you can place the plane vertically or horizontally. When the type of conic sections change, you have to find what it is. You will get extra help at the bottom window.



Help

Classifying conics

page:*CC*

Conics come in three flavours which are most easily seen in the basic examples:

		Cuts the x axis	Cuts the y axis
$\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$	**the ellipse**	$\pm a$	$\pm b$
$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$	**the hyperbola**	$\pm a$	nowhere
$x^2 - 2ay = 0$	**the parabola**	0	0

These cut the x-axis at , $\pm a$, $\pm a$, 0 respectively and the y-axis at $\pm b$, nowhere, 0, respectively.
In particular, if a=b in the ellipse, we have a circle.

$$x^2+y^2 = a^2$$

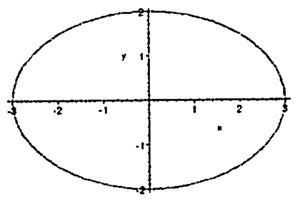
Classifying conics from their graph page:*CC-G*

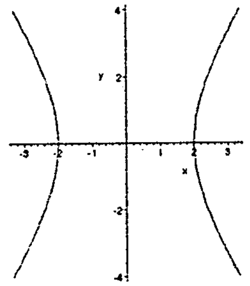
Recognising conic sections from their graph

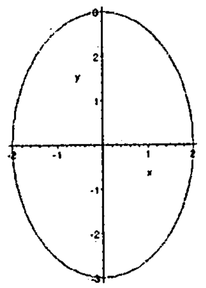
Find which graph best fits the equation

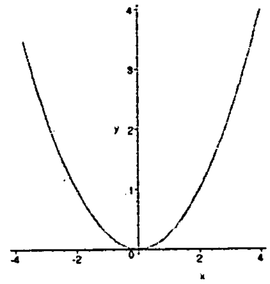
$$\frac{x^2}{4} - \frac{y}{9} = 0$$

Hint | Answer

(1) ○ 

(2) ○ 

(3) ○ 

(4) ○ 

Classifying conics from their equation

page:*CC-E*

Conic sections classification

Classify the following conic sections:

	Ellipse	Hyperbola	Parabola
1. $\frac{x^2}{4} + \frac{y^2}{9} = 1$	○	○	○
2. $\frac{x^2}{4} - \frac{y^2}{9} = 1$	○	○	○
3. $\frac{x^2}{4} - \frac{y}{9} = 0$	○	○	○

Converting quadratic to its standard form

page:***SF***

Given a quadratic in the form

$$\{(x,y) \in \mathbf{R}^2 \mid \alpha x^2+2\beta xy+\gamma y^2+\delta x+\varepsilon y+f=0\}$$

It is possible to convert it to a standard form.

We first eliminate the xy term. This is done by writing the quadratic terms (x^2, xy and y^2) in the following form

$$\begin{bmatrix} x & y \end{bmatrix} \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \mathbf{x}^T A \mathbf{x}$$

where $\mathbf{x}=(x,y)$.

The matrix:

$$A = \begin{bmatrix} \alpha & \beta \\ \beta & \gamma \end{bmatrix}$$

is the equations' *associated matrix*.

Now, if we can find a matrix P in O_2 such that $P^{-1}AP=D$ is diagonal then we can write this as $\mathbf{x}'^T D\mathbf{x}'$, where $\mathbf{x}'=(u,v)$ are new coordinates.

Recall that this amounts to finding the eigenvalues and eigenvectors of the matrix A. The eigenvectors form the columns of P. In other words, we have written our quadratic as

$\lambda_1 u^2+\lambda_2 v^2+ [\delta\ \varepsilon]P\mathbf{x}' +f=0$,

or

$\lambda_1 u^2+\lambda_2 v^2+\delta' u+\varepsilon' v+f=0$,

where λ_1 and λ_2 are the eigenvalues of A and δ', ε' are the appropriate values. Note that in order to obtain a usual orientation of the conic we choose the order of λ_1 and λ_2 to satisfy $1/\lambda_1 > 1/\lambda_2$

Now, we can complete the square in u and v to eliminate the linear terms. If we cannot do this then it is because we have a parabola which has a quadratic term and a linear term only.
In summary, we have an algorithm to simplify a general conic:

1. Find A from α, β and γ.
2. Find the eigenvalues of A
3. It is now possible to classify the conic.
4. Find the eigenvectors of A and form P (you can always choose them to be unit vectors and choose their direction so that P is a rotation matrix).
5. Use P to find the linear terms (they may already be zero in simple examples).
6. Complete the squares to eliminate one or both of the linear terms.
7. Find the standard form and the rotation that would normalise it and even sketch its graph.

Converting quadratic exercise

page:*SF-P*

Put the following conic into standard form, identify it and find the rotation which is needed to rotate the original conic into the standard form.

$$x^2+6xy+y^2=2.$$

1. First find the associated matrix

$$A = \begin{bmatrix} \square & \square \\ \square & \square \end{bmatrix}$$

Check Answer | Hint | Answer

2. Then find the eigenvalues of A: $\square$ and $\square$

Check Answer | Hint | Answer

3. Now write the equation in standard form - use (/) for fractions, for example 1/3, not decimals.

$$\frac{x^2}{\square} \;\square\; \frac{y^2}{\square} = \square$$

Check Answer | Hint | Answer

4. Identify the conic section

(1) ○ ellipse
(2) ○ hyperbola
(3) ○ parabola

5. Find the eigenvectors of P (normalise them to have vector length 1 and choose their direction so that P is a rotation matrix) and write down P:

$$P = \frac{1}{\sqrt{\square}} \begin{bmatrix} \square & \square \\ \square & \square \end{bmatrix}$$

Check Answer | Hint | Answer

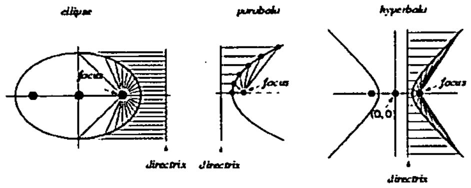
6. What is the angle of the above rotation ?

(1) ○ $\pi/2$
(2) ○ $\pi/3$
(3) ○ $\pi/4$
(4) ○ $\pi/5$

Eccentricity

page:*EC-A*

As mentioned in the focus-directrix property page the eccentricity **e** is the ratio of the distance from the focus, over the horizontal distance from the directrix. Therefore the eccentricity characterises the obtained locus. According to the eccentricity we get different conic sections.



'Play' with the following multiple choice question to see how you can classify conics from their eccentricity. If you don't know just guess and the system will help you. At the end of the activity you will be provided with a summary.

page:*EC-B*

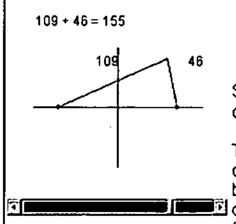
Conic sections classification and eccentricity

Classify the following conic sections from their eccentricity

	Ellipse	Hyperbola	Parabola	Circle
(a) e = 0.172	○	○	○	○
(b) e = 1	○	○	○	○
(c) e = 1.09	○	○	○	○
(d) e = 0	○	○	○	○

Ellipse as a locus

page:*EL-L*



The ellipse can be also characterized as the locus of points for which the sum of the distance from two fixed points (the foci) is constant.

Since the foci of an ellipse are a distance $2ae$ apart, the sum of the distances of any point from the foci is $2a$. (proof)

The interactive activity on the left demonstrates exactly that. Click and drag with your mouse in a circular motion around the foci. The boundary of the ellipse will be drawn and the distances to the foci calculated. Notice that the sum of the distances are always constant (computational rounding errors will cause it not to be always constant but in reality it is always constant). You can change the length of the horizontal axis by simply moving the horizontal scrollbar.

Proof

Choose an arbitrary point P. The sum r+s from the figure is equal to

$$r+s=\frac{2a}{e}$$

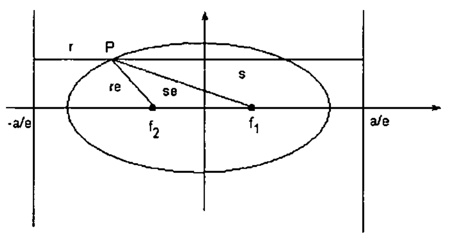
But the distance of any point of an ellipse to the foci is a factor of e times the distance to the corresponding directrix hence

Pf_1 = se and Pf_2 = re

Hence the sum

$re+se = (r+s)e = 2(a/e)e = 2a$

as required.



Eccentricity of an ellipse

page:*EC-E*

Eccentricity is an importand factor of conics as it shows how 'elliptical' the conic section is (i.e. how far away from being a circle that has eccentricity=0)

The parameters a and b of the standard form:

$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} = 1$$

are related to *the eccentricity for an ellipse* by

$$e = \frac{\sqrt{(a^2-b^2)}}{a} \text{ assuming } a > b$$

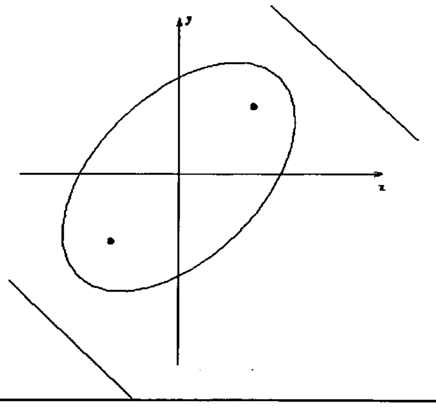
For the hyperbola, the eccentricity is

$$e = \frac{\sqrt{(a^2+b^2)}}{a} \text{ assuming } a > 0$$

The foci (for both cases) are at $(\pm ae, 0)$ and the directrices are given by $x = \pm a/e$.

For the parabola, the focus is at $(0, a/2)$ and the directrix at $y=-a/2$

Example



Content Goals

Part of the XML structure that describes the association between the material in WALLIS as well as their goals that are presented to the student (see Section 5.4.3.1).


```
<goals>
 ...
 <goal dbId="49">
   <goal>
     ...
     <goal dbId="104" type="theory">
       <linktext>
          Go to <link>the geometric definition</link> page
       </linktext>
       <description>
          to read how conic sections arise geometrically
       </description>
      </goal>
      <goal>
       <goal dbId="107" type="theory">
        <description>
            to learn how to classify conic sections
        </description>
        <linktext>
            Go to <link> the types of conics </link> page
        </linktext>
      </goal>
      <goal>
       <goal dbId="109" type="theory">
        <linktext>
            Go to <link> finding standard form  </link> page
          </linktext>
          <description>
            to read the theory behind converting a quadratic equation
             to its standard form
          </description>
      </goal>
      ...
```

```
      <goal dbId="111" type="example">
         <linktext>
            Go to <link> this self-practice</link> page
         </linktext><description>
            to practice converting a quadratic to its standard form
         </description>
         <goal dbId="1101" type="part"/>
         <goal dbId="1102" type="part"/>
         <goal dbId="1103" type="part"/>
         <goal dbId="1104" type="part"/>
         <goal dbId="1105" type="part"/>
         <goal dbId="1106" type="part">
            <script>
              return docFrame.checkAttempt(s);
            </script>
         </goal>
      </goal>
   </goal>

</goals>


</exercise>
```

Appendix C

Materials for Studies

C.1 Student self-reports study

This section presents the materials used in the study described in Section 6.2 for collecting students' self-reports on their emotions and motivation. The document on pages 214-215 was used to present the goals of the research to students and the task they had to complete. The questionnaire that was given to them in an A3 paper is shown on page 216. Pages 217-219 present the pre-test which was given to them before interacting with the system.

Please spend some time reading the following

Introduction

The purpose of this study is to investigate further the use of Interactive Learning Environments such as WaLLiS[1] and how do student interact with such systems and how they feel during their interaction. For that I rely a lot on your feedback, openness and frankness. As I am interested in the actual use of the system I am asking you to interact as you would normally do in any other kind of learning activity (for example in your own time and pace, keeping notes to revise etc). On the other hand, there are some restrictions that I am obliged to impose due to the nature of my research. Therefore,

- please do not 'collaborate' or consult other students
- please stick to the material that are available in WaLLiS and do not refer to a book or other online material

I appreciate that some of these restrictions are unnatural (and perhaps anti-pedagogical) but as soon as the study is over you can go over the material again in combination with your books or your fellow students. But please do not do that during the study as we want to measure the impact of the system and the way it is used. Afterwards, you will also be given the opportunity to cover anything you haven't understood.

Description of task

Set some time aside when you would study for the GC course and instead login and try to learn from WaLLiS.

You can access[2] the system at http://student.maths.ed.ac.uk/wallis.
Remember: you have to register first (as shown in class).

Your main goal is to acquire the knowledge of these pages. You have a week to do that. Imagine the system as a resource for material that you have to learn. WaLLiS will try to help you as much as possible with suggestions and hints in the exercises. Note that apart from the fact that these materials are part of the curriculum and you would have to learn them anyway, they will be needed for a part of next week's assessment and probably seem useful at the final exams.

---

[1] WaLLiS got its name from a mathematician who in 1655 published a treatise on conics.
[2] Unfortunately, for the purposes of this study you can only use Internet Explorer. You can use your own computer if you have (or are willing to install) the Java Plugin. Otherwise, you can always use the university labs where WaLLiS works fine.

Note that your interaction with the system will only be known to me and **not your lecturer** and **the way you interact or work with the system will not have a negative impact** in terms of marks or the lecturer's or tutors' opinions on you. Therefore feel free to explore and work with the system as you like.

At some point next week we will arrange an interview to go over your interaction with the system and discuss aspects that pertain to my research. Although your interaction will be recorded, it is hard for people to remember how they felt when they were doing something. **Therefore, you are provided with the attached table as a guide for keeping notes that will help you remember interesting moments of your interaction with the system**: *moments that relate to your motivation, feelings or your requests for help or hints.*

So ... have a look at the table in advance to get an idea of what I am looking for you to remember (aspects like your **confidence, effort, hesitation, frustration** and also your requests for **help and hints).** To help you reference the pages they have IDs (like page:SF-E) and are also printed at the back of the table. Note that if you feel you want to keep a note on something interesting that you noticed which is not included on the table, there is more space at the back page where you can write additional comments (for example 'part 2 of page SF-E: I found it very easy so...' or 'page:SF I: was distracted so took me some time to complete' etc.)

<u>You do not have to concentrate on completing the table during the interaction</u>. If it is distracting you can leave it aside and immediately when you finish or pause (even if you have to stop to continue some other time) complete the table with as much information as you can. If you return back to the system later you could just go through the table again and add to it.

If you are not sure about something or if you have technical problems feel free to contact me anytime at m.mavrikis@ed.ac.uk or if it is urgent feel free to contact me at xxxxxxxxxx.

The next step (after your interaction with the system) is that I will send you a reminder to arrange an interview. You will have to bring the table with you in order to remind you the issues and problems you faced and to help me focus our discussion. You will also take a very quick test to gauge how useful the system was for your learning. Results and feedback as well as further help (if you require it) will be provided after the interview.

Thanks in advance for all your help,

Manolis Mavrikis

Use the following table to help you remember at which parts you felt as described in the first column. Apart from keeping a note of the page, its part (e.g. SF-P:2) and the time, try to keep a brief note of the situation and why you felt like that.

Effort **Lows-highs** times, places you spend the most or least amount of effort (or mental energy) to work with the material or exercise	
Hesitation points where you think you particularly hesitated or were reluctant to provide an answer.	
Confidence **Lows-highs** your belief that you know the material or how to perform a certain step. Try to think of it regardless of the answer of the system and keep a note to remember how you felt after	
Confused didn't know what to do or what page to study. What did you do to get away from this ?	
Bored demotivated, weary or lacking of interest ? why? what did you do for that?	
Interested in the sense that this part increased your desire to learn and perhaps curious to look it up further	
Frustrated/Annoyed times that you felt that you are disappointingly unsuccessful why was that that?	
Contented/'happy' times when you think you were enjoying or satisfied with the situation	
Help/Hint Requests write down particular interesting instances where you asked for help or hints and how you felt about the help provided	

turn page to see the web-pages and to keep any other notes

GIC Questionnaire 1 Feb06

Matric: ....

General questions

1) Do you have a rough idea of what conic sections are ?

2) Were you taught conic sections before and where ?

3) What would you say is your knowledge on conics sections ?

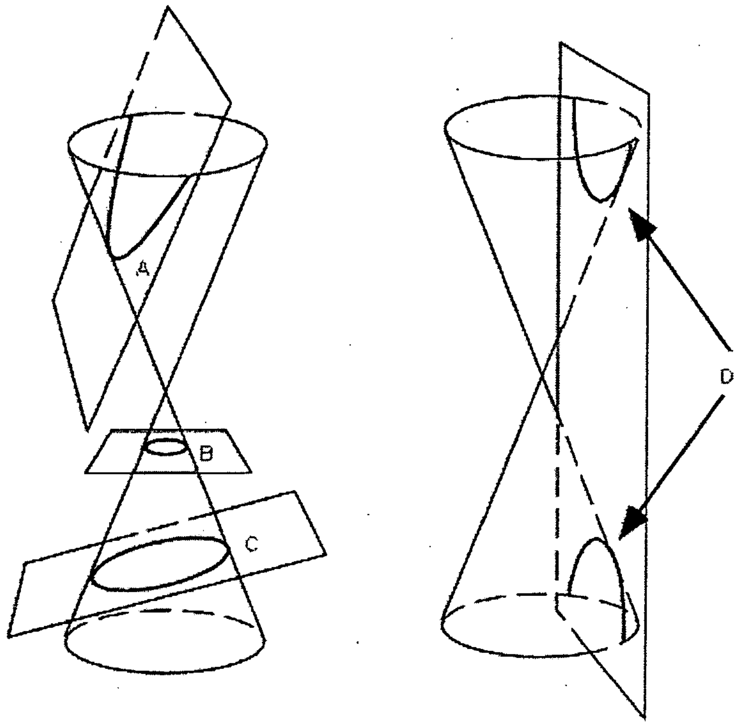
4) How confident do you feel on your knowledge of conics sections ?

5) Usually there is a geometric approach/definition and an algebraic one. Do you remember this fact? Do you feel familiar with both, none or one of them ?

if you answered yes in any of the above then continue to the next page

Questions

1) In the geometric definition conic sections are the curves which are generated by the intersections of a plane with one or two nappes (pieces) of a cone. As follows:



These curves have distinct names. If you remember some of them write them here:

A: ....

B: ....

C: ....

D: ....

2) In the algebraic definition these curves are associated with a quadratic equation

$$\{(x,y) \in \mathbf{R}^2 \mid \alpha x^2 + 2\beta xy + \gamma y^2 + \delta x + \varepsilon y + f = 0\}$$

where you aware of this (we don't mean the exact details but the fact that there is a quadratic expression behind these curves)

continue to the next page

3) It is a fact that ellipses, circles and hyperbolas can be written in a simpler form called **standard form** that looks like the following :

$$\frac{x^2}{a^2} \pm \frac{y^2}{b^2} = 1$$

Have you ever seen this fact before and if yes did you remember that ?

Please don't guess in the following question if you don't know then just say so

4) Can you find which of the following equations represents a circle, an ellipse, a hyperbola or a parabola ?

$$\frac{x^2}{4} + \frac{y^2}{9} = 1$$

$$x^2 + y^2 = 9$$

$$\frac{x^2}{a^2} - \frac{y^2}{b^2} = 1$$

$$y^2 - 2ax = 0$$

C.2 Tutor diagnosis study

This section presents the environment used for the study in Section 6.6.1 as well as a sample student-tutor dialogue. The description of the environment below is adapted from Porayska-Pomsta et al. (2008) and Andres et al. (2005a).

C.2.1 The environment

Two interfaces were used in the study, one for use by tutors and one by students. The data were recorded based on the same interaction capture agent developed for WALLIS (slightly modified to accommodate the needs of this study). The two interfaces differed in a number of respects.

The student interface, shown in Figure C.1, comprised:

Theory frame: (top frame) in which the students referred to background material related to the exercises, if directed to do so by the tutor;

Text and maths editor: (bottom left frame) which permitted writing of mathematical formulae (middle box) as well as any text around the formula (in upper and lower text boxes). The maths editor[1] was equipped with maths templates from which the students could select a set of pre-defined formulae appropriate for a given exercise.

History of the interaction: (bottom right frame), indicating student and tutor interaction, through which the student could scroll at any time during a session.

To enable the tutor to indicate the situational factors, and the values that they deemed relevant to their feedback, a frame with a set of pre-defined factors was provided. The situational factors selection tool was purpose built. The tutor's screen, shown in Figures C.2, comprised:

Text and maths editor: (top left) as in the student interface;

[1] a slightly modified version of WebEq (http://www.dessci.com/en/products/webeq)

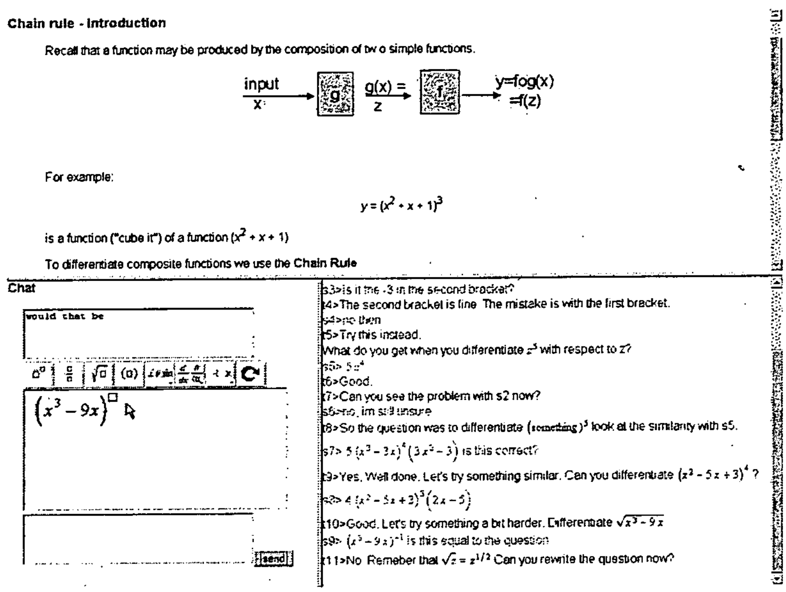


Figure C.1: Student chat interface

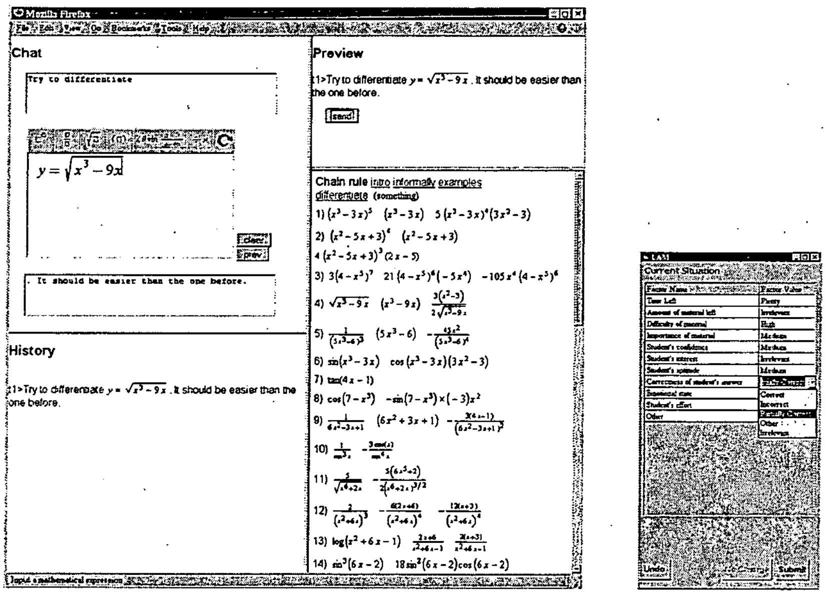


Figure C.2: Tutors chat interface and situational factors selection tool. The two factors defined by the tutors are *Emotional state* and *Student's effort*.

History of the interaction: (bottom left), indicating student and tutor interaction, through which the tutor could scroll at any time during a session;

Preview frame: (top right) for previewing the feedback typed before sending it to the student;

Exercise frame: (bottom right) in which pre-specified chain rule exercises were provided for use during the session.

Situational factors selection tool: which the tutor used to identify the situational factors and the values relevant to their feedback.

In addition to the screen through which the tutor could interact with the student and select situational factors, tutors were provided with a second screen that was connected to the students' computer and through which they could observe all student actions at all times during an interaction. Students were informed of this explicitly before every session.

C.2.2 A sample dialogue

Logfile: tut3 - std4-v1

Factors

Thu Nov 11 15:07:12 GMT 2004	tm1:	Hello. Do you remember the chain rule?
Thu Nov 11 15:07:28 GMT 2004	sm1:	i think so
Thu Nov 11 15:08:13 GMT 2004	tm2:	Let's try an example first. Try to differentiate the following $(x^3-3x)^5$
Thu Nov 11 15:09:11 GMT 2004	sm2:	$5(x^3-3x)^4*(2x^2-3)$
Thu Nov 11 15:09:33 GMT 2004	sm3:	do i need to go further than that?
Thu Nov 11 15:10:03 GMT 2004	tm3:	That's good. You seem to know what's going on. I think that will do for that one. Try $3(4-x^5)^7$
Thu Nov 11 15:11:09 GMT 2004	sm4:	$21(4-x^5)^6x(-4x^4)$
Thu Nov 11 15:11:52 GMT 2004	tm4:	OK. That's not quite right. Can you spot the error?
Thu Nov 11 15:12:27 GMT 2004	sm5:	sorry $21(4-x^5)^6x(-5x^4)$
Thu Nov 11 15:13:00 GMT 2004	tm5:	That's it. Can you re-arrange it to make it look neater?
Thu Nov 11 15:13:22 GMT 2004	sm6:	do you mean multiply the two together?
Thu Nov 11 15:13:55 GMT 2004	tm6:	I mean write it in a shorter neater way perhaps by multiplying certain parts
Thu Nov 11 15:13:55 GMT 2004	sm7:	no i dont think so. Im not exactly remotely competent
Thu Nov 11 15:15:19 GMT 2004	tm7:	You're doing pretty well up to now. What I mean is to bring the numerical parts to the front
Thu Nov 11 15:15:59	sm8:	$84-21x^{11}*(-5x^4)$

GMT 2004		
Thu Nov 11 15:15:59 GMT 2004	sm9:	$84 - 21x^{30} * (-5x^4)$
Thu Nov 11 15:16:38 GMT 2004	tm8:	Why did you do that?
Thu Nov 11 15:16:59 GMT 2004	sm10:	because i thought you mulitplied the powers when they were bracketed
Thu Nov 11 15:18:11 GMT 2004	tm9:	Sorry, a bit of a technical problem at this end. To continue: what I mean is to multiply together the numerical factors (-5) and (21)
Thu Nov 11 15:19:30 GMT 2004	sm11:	$-105(4-x^5)^6 * x^4$
Thu Nov 11 15:20:16 GMT 2004	tm10:	That's right. Exactly. In fact it's conventional to write it as $-105x^4(4-x^5)^6$
Thu Nov 11 15:20:26 GMT 2004	sm12:	i see. i shouldve known that really
Thu Nov 11 15:21:38 GMT 2004	tm11:	That's OK. Now let's try $\frac{1}{6x^2-3x+1}$
Thu Nov 11 15:23:04 GMT 2004	sm13:	$\frac{1}{6x^2-3x+1}$ $6x^{-2} - 3x^{-1} + 1$
Thu Nov 11 15:23:43 GMT 2004	tm12:	Do you mean the whole thing is the derivative?
Thu Nov 11 15:24:00 GMT 2004	sm14:	sorry. im being a fool.
Thu Nov 11 15:24:40 GMT 2004	tm13:	Try to think what the inner layer is
Thu Nov 11 15:25:18 GMT 2004	sm15:	should the derivative be: $12x^{-1} + 3x^{-2}$
Thu Nov 11 15:26:18 GMT 2004	tm14:	Remember that $\frac{1}{a+b}$ is not equal to $\frac{1}{a} + \frac{1}{b}$
Thu Nov 11 15:26:54 GMT 2004	sm16:	i really dont know what i should be doing with this equation
Thu Nov 11 15:27:24 GMT 2004	tm15:	So $\frac{1}{a}$ equals a^{-1}

Thu Nov 11 15:28:23 GMT 2004	sm17:	yes. i dont see what i should be doing.
Thu Nov 11 15:29:08 GMT 2004	tm16:	Lets' split the problem into finding the inner layer and the outer layer. The problem is similar to the first one. Cna you see what the outer layer should be?
Thu Nov 11 15:29:44 GMT 2004	sm18:	no
Thu Nov 11 15:30:17 GMT 2004	tm17:	What about something like $\frac{1}{z}$
Thu Nov 11 15:33:01 GMT 2004	sm19:	$(x)^{-1}$
Thu Nov 11 15:33:57 GMT 2004	tm18:	Try thinking about $z = 6x^2 - 3x + 1$
Thu Nov 11 15:32:12 GMT 2004	sm20:	Do you want me to factorise it?
Thu Nov 11 15:33:29 GMT 2004	tm19:	Not really. I mean consider $f(x) = z^{-1}$ and$z = 6x^2 - 3x + 1$ Just as we did in the first problem
Thu Nov 11 15:34:11 GMT 2004	sm21:	so i should find its derivative?
Thu Nov 11 15:36:02 GMT 2004	tm20:	After that technical hitch let's try another problem. Think about $\frac{2}{(x^2+6x)^3}$
Thu Nov 11 15:37:49 GMT 2004	sm22:	if $\frac{1}{x} = x^{-1}$does $\frac{2}{x} = x^{-2}$?
Thu Nov 11 15:37:49 GMT 2004	sm23:	if
Thu Nov 11 15:38:15 GMT 2004	tm21:	No. It's $2x^{-1}$
Thu Nov 11 15:39:57 GMT 2004	sm24:	ok. so can i re-write it: $(2(x^2+6x)^3)^{-1}$?
Thu Nov 11 15:40:21 GMT 2004	sm25:	and then take it from there?
Thu Nov 11 15:40:32 GMT 2004	tm22:	Nearly. The 2 should not be inside the outer bracket

Thu Nov 11 15:40:54 GMT 2004	tm23:	You also need to apply the power laws to rewrite the expression
Thu Nov 11 15:40:54 GMT 2004	sm26:	$2((x^2+6x)^3)^{-1}$
Thu Nov 11 15:41:27 GMT 2004	sm27:	$2(x^2+6x)^{-3}$
Thu Jan 01 01:00:00 CET 1970	sm28:	
Thu Nov 11 15:42:20 GMT 2004	tm24:	That's right. Now can you apply the chain rule to that?
Thu Nov 11 15:42:22 GMT 2004	tm25:	
Thu Nov 11 15:43:24 GMT 2004	sm29:	$-6(x^2+6x)^{-2}(2x+6)$
Thu Nov 11 15:43:42 GMT 2004	tm26:	That's right. Can you simplify that one too?
Thu Nov 11 15:44:33 GMT 2004	sm30:	$-12(x^2+6x)^{-2}-36$
Thu Nov 11 15:44:50 GMT 2004	sm31:	$-12x(x^2+6x)^{-2}-36$
Thu Nov 11 15:45:18 GMT 2004	tm27:	Not quite. Where did the -36 come from?
Thu Nov 11 15:45:31 GMT 2004	sm32:	-6*6
Thu Nov 11 15:46:40 GMT 2004	tm28:	I don't think I meant to multiply the brackets out (even so, should the 36 needs to be multiplied by the expression in the bracket). I meant to collect terms and take out common numerical factors.
Thu Nov 11 15:47:01 GMT 2004	sm33:	ok ill try
Thu Nov 11 15:48:04 GMT 2004	tm29:	I think we need to stop now. What I meant was $-\frac{12(x+3)}{(x^2+6x)^4}$
Thu Nov 11 15:48:24 GMT 2004	sm34:	thats the answer???!!

Thu Nov 11 15:48:41 GMT 2004	sm35:	i would never have got there
Thu Nov 11 15:49:14 GMT 2004	tm30:	I'm afraid so! I'm sure you would... Thanks for your help Ben.

Appendix D

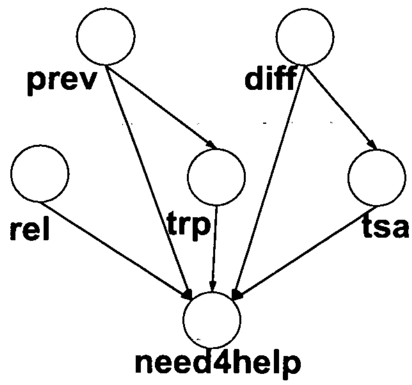
Machine Learning Outputs

D.1 Bayesian Networks

D.1.1 Need for help

This section presents the different models derived using WEKA as described in Section 5.3. The structure and accuracy of the single model is presented first followed by the 9 individual models. The probability tables of the ones with simpler structure are also shown. The Bayesian models are learned employing the merged dataset of `GIC03` and `GIC04` using *HillClimbing* or *ICS* (see Section 3.4.3). For the HillClimbing two further choises were possible, initiate them as Naive Bayesian Networks, that is a network with an arrow from the classifier node to each other node, or present to the algorithm an empty network as initial structure. From the above choices, for every model, whichever optimised the 10-fold cross validation and had higher Kappa and less false negatives was chosen for implementation in JavaBayes. This was tested against the test dataset `GIC05`.

In addition, comparisons against the accuracy of decision trees (using *J4.8*) and *logistic regression* are provided. Section 3.4.6.2 explained the reasons that global pre-discretization, which takes into account properties of the domain, is often preferred. However, better accuracies can be achieved some times when discretization is performed based on the attribute properties in the dataset. For completeness this possibility was tested as well. For each model the more accurate between the pre-descritized dataset and the weka-discretised one is tested against a test dataset. Overall, logistic regression has the best accuracy with the accuracy of bayesian networks being very close, often matching it.



	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	67.640	66.5169	67.640	64.045	65.842	71.910	68.764	69.887
Kappa	0.317	0.318	0.349	0.234	0.297	0.421	0.350	0.377
T	0.743	0.765	0.701	0.719	0.708	0.736	0.745	0.750
F	0.605	0.564	0.647	0.5	0.589	0.650	0.597	0.621

Figure D.1: Srtucture of Bayesian network and accuracy of single model for predicting need for help.

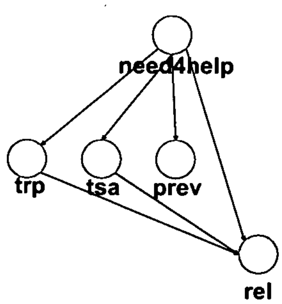


Figure D.2: Bayesian network for predicting the need for help in GD-A

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	69.663	69.841	69.841	69.662	68.254	73.034	64.045	70.635
Kappa	0.317	0.401	0.36	0.276	0.33	0.432	0.226	0.397
T	0.773	0.694	0.763	0.787	0.747	0.782	0.719	0.748
F	0.542	0.703	0.587	0.471	0.574	0.647	0.5	0.648

TRUE	FALSE
0.595	0.405

Probability Distribution Table for trp					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.081	0.279	0.46	0.135	0.045
FALSE	0.092	0.142	0.324	0.221	0.221

Probability Distribution Table for tsa					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.081	0.243	0.334	0.243	0.099
FALSE	0.09	0.273	0.299	0.221	0.117

Probability Distribution Table for prev					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.063	0.045	0.46	0.405	0.027
FALSE	0.195	0.169	0.403	0.22	0.013

Probability Distribution Table for rel				
need4help	tsa	trp	TRUE	FALSE
TRUE	V.LOW	V.LOW	0.75	0.25
TRUE	V.LOW	LOW	0.75	0.25
TRUE	V.LOW	MEDIUM	0.25	0.75
TRUE	V.LOW	HIGH	0.75	0.25
TRUE	V.LOW	V.HIGH	0.5	0.5
TRUE	LOW	V.LOW	0.5	0.5
TRUE	LOW	LOW	0.375	0.625
TRUE	LOW	MEDIUM	0.611111111	0.388889
TRUE	LOW	HIGH	0.25	0.75
TRUE	LOW	V.HIGH	0.25	0.75
TRUE	MEDIUM	V.LOW	0.833333333	0.166667
TRUE	MEDIUM	LOW	0.357142857	0.642857
TRUE	MEDIUM	MEDIUM	0.785714286	0.214286
TRUE	MEDIUM	HIGH	0.875	0.125
TRUE	MEDIUM	V.HIGH	0.75	0.25
TRUE	HIGH	V.LOW	0.25	0.75
TRUE	HIGH	LOW	0.9	0.1
TRUE	HIGH	MEDIUM	0.611111111	0.388889
TRUE	HIGH	HIGH	0.5	0.5
TRUE	HIGH	V.HIGH	0.5	0.5
TRUE	V.HIGH	V.LOW	0.5	0.5
TRUE	V.HIGH	LOW	0.25	0.75
TRUE	V.HIGH	MEDIUM	0.166666667	0.833333
TRUE	V.HIGH	HIGH	0.166666667	0.833333
TRUE	V.HIGH	V.HIGH	0.5	0.5
FALSE	V.LOW	V.LOW	0.5	0.5
FALSE	V.LOW	LOW	0.5	0.5
FALSE	V.LOW	MEDIUM	0.166666667	0.833333
FALSE	V.LOW	HIGH	0.25	0.75
FALSE	V.LOW	V.HIGH	0.5	0.5
FALSE	LOW	V.LOW	0.166666667	0.833333
FALSE	LOW	LOW	0.75	0.25
FALSE	LOW	MEDIUM	0.071428571	0.928571
FALSE	LOW	HIGH	0.5	0.5
FALSE	LOW	V.HIGH	0.75	0.25
FALSE	MEDIUM	V.LOW	0.5	0.5
FALSE	MEDIUM	LOW	0.5	0.5
FALSE	MEDIUM	MEDIUM	0.75	0.25
FALSE	MEDIUM	HIGH	0.833333333	0.166667
FALSE	MEDIUM	V.HIGH	0.9	0.1
FALSE	HIGH	V.LOW	0.5	0.5
FALSE	HIGH	LOW	0.5	0.5
FALSE	HIGH	MEDIUM	0.625	0.375
FALSE	HIGH	HIGH	0.7	0.3
FALSE	HIGH	V.HIGH	0.25	0.75
FALSE	V.HIGH	V.LOW	0.75	0.25
FALSE	V.HIGH	LOW	0.5	0.5
FALSE	V.HIGH	MEDIUM	0.5	0.5
FALSE	V.HIGH	HIGH	0.75	0.25
FALSE	V.HIGH	V.HIGH	0.833333333	0.166667

Figure D.3: Accuracy and probability tables for GD-A

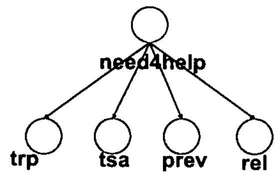


Figure D.4: Bayesian network for predicting the need for help in CC-G

TRUE	FALSE
0.561	0.439

Probability Distribution Table for trp

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.086	0.2	0.39	0.219	0.105
FALSE	0.084	0.253	0.422	0.108	0.133

Probability Distribution Table for tsa

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.029	0.181	0.314	0.333	0.143
FALSE	0.157	0.349	0.325	0.109	0.06

Probability Distribution Table for prev

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.086	0.048	0.505	0.333	0.028
FALSE	0.157	0.157	0.349	0.325	0.012

Probability Distribution Table for rel

need4help	TRUE	FALSE
TRUE	0.559	0.338
FALSE	0.441	0.662

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	68.539	68.254	68.539	65.269	69.843	74.157	61.798	70.9492
Kappa	0.361	0.37	0.354	0.266	0.389	0.476	0.224	0.404
T	0.72	0.714	0.731	0.726	0.729	0.768	0.66	0.75
F	0.641	0.643	0.622	0.523	0.661	0.709	0.564	0.6508

Figure D.5: Accuracy and probability tables for predicting the need for help in CC-G

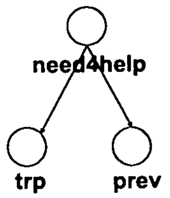


Figure D.6: Bayesian network for predicting the need for help in CC-E

TRUE	FALSE
0.561	0.439

Probability Distribution Table for trp

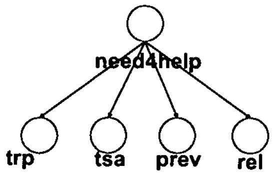
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.048	0.124	0.409	0.276	0.143
FALSE	0.133	0.349	0.398	0.036	0.084

Probability Distribution Table for prev

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.0667	0.028	0.486	0.39	0.029
FALSE	0.181	0.181	0.373	0.253	0.012

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	71.91	69.841	67.46	64.045	65.873	70.786	66.667	69.841
Kappa	0.421	0.387	0.319	0.257	0.329	0.403	0.332	0.389
T	0.762	0.732	0.732	0.698	0.661	0.745	0.691	0.729
F	0.658	0.653	0.586	0.894	0.656	0.658	0.638	0.661

Figure D.7: Accuracy and probability tables for CC-E



TRUE	FALSE
0.561	0.439

Probability Distribution Table for trp					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.066	0.181	0.448	0.2	0.105
FALSE	0.108	0.277	0.349	0.133	0.133

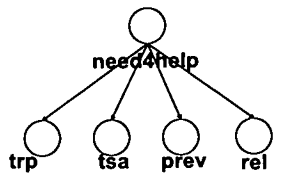
Probability Distribution Table for tsa					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.01	0.2	0.333	0.314	0.143
FALSE	0.181	0.325	0.301	0.133	0.06

Probability Distribution Table for prev					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.105	0.048	0.486	0.333	0.028
FALSE	0.133	0.157	0.373	0.325	0.012

Probability Distribution Table for rel		
need4help	TRUE	FALSE
TRUE	0.5	0.337
FALSE	0.5	0.663

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	69.663	67.46	65.873	56.179	56.349	69.46	65.168	67.46
Kappa	0.375	0.336	0.294	0.087	0.131	0.393	0.286	0.336
T	0.743	0.725	0.726	0.642	0.56	0.721	0.699	0.725
F	0.63	0.602	0.547	0.435	0.56	0.672	0.587	0.602

Figure D.8: Bayesian network for SF-P-AM



TRUE	FALSE
0.572	0.428

Probability Distribution Table for trp

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.028	0.196	0.346	0.252	0.178
FALSE	0.161	0.259	0.481	0.062	0.037

Probability Distribution Table for tsa

need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.047	0.234	0.271	0.308	0.14
FALSE	0.136	0.284	0.382	0.136	0.062

Probability Distribution Table for prev

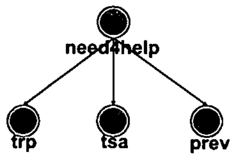
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.028	0.066	0.551	0.327	0.028
FALSE	0.235	0.136	0.284	0.333	0.012

Probability Distribution Table for rel

need4help	TRUE	FALSE
TRUE	0.433	0.526
FALSE	0.567	0.474

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	69.662	64.286	63.492	58.427	59.524	70.786	65.079	68.254
Kappa	0.378	0.261	0.315	0.118	0.272	0.399	0.242	0.33
T	0.738	0.713	0.603	0.673	0.622	0.75	0.732	0.75
F	0.64	0.526	0.662	0.431	0.564	0.649	0.5	0.565

Figure D.9: Bayesian network for SF-P-SF-C



TRUE	FALSE
0.506	0.494

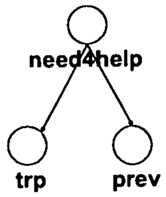
Probability Distribution Table for trp					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.031	0.179	0.474	0.179	0.137
FALSE	0.183	0.183	0.376	0.118	0.14

Probability Distribution Table for tsa					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.095	0.053	0.431	0.368	0.053
FALSE	0.161	0.226	0.398	0.161	0.054

Probability Distribution Table for prev					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.053	0.053	0.474	0.389	0.031
FALSE	0.183	0.14	0.398	0.269	0.01

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	65.873	64.045	64.045	53.932	60.317	67.415	66.292	65.873
Kappa	0.322	0.28	0.278	0.074	0.211	0.347	0.324	0.32
T	0.715	0.667	0.692	0.631	0.671	0.695	0.694	0.699
F	0.574	0.61	0.568	0.388	0.5	0.651	0.625	0.606

Figure D.10: Bayesian network for SF-P-SF-C



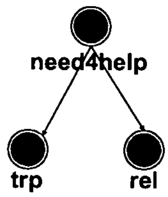
TRUE	FALSE
0.561	0.439

Probability Distribution Table for trp					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.048	0.086	0.467	0.257	0.142
FALSE	0.157	0.398	0.253	0.084	0.108

Probability Distribution Table for prev					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.01	0.086	0.448	0.428	0.028
FALSE	0.253	0.108	0.422	0.205	0.012

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	64.044	65.873	68.539	64.044	54.761	71.91	62.921	72.222
Kappa	0.266	0.322	0.353	0.255	0.101	0.421	0.252	0.4453
T	0.692	0.719	0.741	0.714	0.627	0.762	0.66	0.733
F	0.568	0.566	0.6	0.515	0.424	0.658	0.593	0.711

Figure D.11: Bayesian network for SF-P-SF-EG1



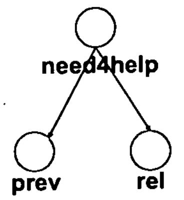
TRUE	FALSE
0.661	0.339

Probability Distribution Table for trp					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.057	0.203	0.333	0.236	0.171
FALSE	0.138	0.263	0.538	0.046	0.015

Probability Distribution Table for *rel*					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.041	0.073	0.138	0.74	0.008
FALSE	0.262	0.385	0.323	0.015	0.015

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	71.91	69.048	69.662	64.044	67.46	73.033	65.079	65.168
Kappa	0.387	0.394	0.356	0.255	0.371	0.442	0.307	0.253
T	0.783	0.742	0.757	0.714	0.725	0.657	0.656	0.726
F	0.603	0.614	0.597	0.515	0.602	0.778	0.645	0.523

Figure D.12: Bayesian network for SF-P-SF-ROT



TRUE	FALSE
0.528	0.472

Probability Distribution Table for prev					
need4help	V.LOW	LOW	MEDIUM	HIGH	V.HIGH
TRUE	0.03	0.051	0.515	0.374	0.03
FALSE	0.214	0.146	0.348	0.281	0.011

Probability Distribution Table for rel		
need4help	TRUE	FALSE
TRUE	0.573	0.36
FALSE	0.427	0.64

	Bayesian		J48			Logistic		
	Cross	TestSet	Cont	Discr	Best with test	Cont	Discr	Best with test
Class	70.786	69.841	73.033	70.768	65.079	73.033	71.91	69.841
Kappa	0.415	0.413	0.444	0.419	0.281	0.459	0.429	0.413
T	0.723	0.712	0.774	0.711	0.577	0.745	0.762	0.712
F	0.69	0.683	0.667	0.705	0.703	0.714	0.658	0.683

Figure D.13: Bayesian network for SF-P-SF-EG2

D.2 Decision Trees - Student Pespective

D.2.1 Confidence

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     temp_try_wallis_need4hint
Instances:    538
Attributes:   action, time, need4hint, diff, std_ability,
              hist_vector {answer_incorrect, answer_partial},
              conf_change

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

action = reading_theory_example
|   difficult <= 2: no_report (89.0)
|   difficult > 2
|   |   std_ability <= 3: decrease (27.0/2.0)
|   |   std_ability > 3: no_report (27.0)
action = comfirm_answer_incorrect
|   answer_incorrect <= 2
|   |   hints_info <= 0
|   |   |   std_ability <= 2
|   |   |   |   difficult <= 1: no_report (10.0)
|   |   |   |   difficult > 1: decrease (17.0)
|   |   |   std_ability > 2
|   |   |   |   need4hint = yes: increase (6.0)
|   |   |   |   need4hint = no: no_report (12.0)
|   |   hints_info > 0: decrease (28.0/2.0)
|   answer_incorrect > 2: decrease (11.0/2.0)
action = comfirm_answer_partial
|   answer_partial <= 0: increase (27.0)
|   answer_partial > 0
|   |   answer_incorrect <= 2: decrease (9.0)
|   |   answer_incorrect > 2: increase (7.0)
action = comfirm_answer_correct: no_report (29.0)
action = request_sol_partial: no_report (28.0/4.0)
action = request_hint_incorrect
|   answer_partial <= 0
|   |   need4hint = yes
|   |   |   answer_incorrect <= 1: no_report (9.0)
|   |   |   answer_incorrect > 1: decrease (19.0/1.0)
|   |   need4hint = no: decrease (39.0)
```

```
|   answer_partial > 0
|   |   time = Low: decrease (26.0/5.0)
|   |   time = Medium: increase (13.0/1.0)
|   |   time = High: increase (10.0/1.0)
action = request_hint_partial
|   diff <= 2: increase (19.0/5.0)
|   diff > 2: decrease (22.0/3.0)
action = read_statement
|   std_ability <= 1
|   |   difficult <= 1: no_report (5.0/1.0)
|   |   difficult > 1: decrease (22.0/3.0)
|   std_ability > 1: no_report (27.0/3.0)

Number of Leaves  :     24

Size of the tree :      41


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         505               93.8662 %
Incorrectly Classified Instances        33                6.1338 %
Kappa statistic                          0.9003
Mean absolute error                      0.0525
Root mean squared error                  0.1619
Relative absolute error                 17.1308 %
Root relative squared error             41.4005 %
Total Number of Instances              538

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.916     0.028      0.966     0.916     0.94     no_report
  0.957     0.055      0.918     0.957     0.937    decrease
  0.987     0.015      0.915     0.987     0.949    increase
  0         0          0         0         0        extreme_decrease

=== Confusion Matrix ===

   a   b   c   d   <-- classified as
 228  15   6   0 |   a = no_report
   8 202   1   0 |   b = decrease
   0   1  75   0 |   c = increase
   0   2   0   0 |   d = extreme_decrease
```

D.2.2 Effort

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -C 0.25 -M 2
Relation:     effort_wallis3
Instances:    369
Attributes:   10
              action
              difficult
              std_ability
              answer_incorrect
              answer_partial
              hints_info
              hints_partial
              time
              pknow
              class
Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 pruned tree
------------------

action = comfirm_answer_partial
|   time = low
|   |   answer_partial <= 0: no_report (39.32/0.32)
|   |   answer_partial > 0: increase (4.03/0.03)
|   time = medium: increase (18.15/2.15)
|   time = high: increase (20.16/4.16)
action = comfirm_answer_incorrect
|   answer_incorrect <= 0: no_report (48.39/12.39)
|   answer_incorrect > 0
|   |   time = low: decrease (8.07/3.07)
|   |   time = medium: increase (10.08/0.08)
|   |   time = high: increase (9.07/0.07)
action = hint_info
|   answer_incorrect <= 1: no_report (48.39/6.39)
|   answer_incorrect > 1
|   |   difficult <= 1: decrease (14.11/0.11)
|   |   difficult > 1
|   |   |   difficult <= 2
|   |   |   |   time = low: decrease (4.03/0.03)
|   |   |   |   time = medium: no_report (4.03/0.03)
|   |   |   |   time = high: no_report (0.0)
|   |   |   difficult > 2: no_report (10.08/0.08)
action = hints_partial
|   hints_partial <= 1: no_report (16.13/0.13)
|   hints_partial > 1: increase (8.07/0.07)
```

```
action = comfirm_answer_correct
|   hints_info <= 1: increase (27.22/0.22)
|   hints_info > 1: no_report (27.22/0.22)
action = prompt_exit_theory_example: no_report (19.16/1.16)
action = prompt_exit_other
|   pknow = yes: no_report (7.06/1.06)
|   pknow = no: decrease (7.06/1.06)
|   pknow = _: no_report (0.0)
action = prompt_stay
|   pknow = yes: no_report (11.09/3.09)
|   pknow = no: increase (8.07/0.07)
|   pknow = _: increase (0.0)

Number of Leaves  :     25

Size of the tree :      38


Time taken to build model: 0.16 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         329               89.1599 %
Incorrectly Classified Instances        40               10.8401 %
Kappa statistic                          0.799
Mean absolute error                      0.0561
Root mean squared error                  0.1725
Relative absolute error                 29.6923 %
Root relative squared error             56.3533 %
Total Number of Instances              369

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.891     0.039      0.907     0.891     0.899    increase
  0.963     0.168      0.888     0.963     0.924    no_report
  0.595     0.012      0.862     0.595     0.704    decrease
  0         0          0         0         0        366.0
  0         0          0         0         0        213.0
  0         0          0         0         0        153.0

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
  98  10   2   0   0   0 |   a = increase
   6 206   2   0   0   0 |   b = no_report
   4  13  25   0   0   0 |   c = decrease
   0   1   0   0   0   0 |   d = 366.0
   0   1   0   0   0   0 |   e = 213.0
```

```
  0   1   0   0   0   0 |    f = 153.0



=== Run information ===

Scheme:          weka.classifiers.trees.J48 -U -M 2
Relation:        effort_wallis2
Instances:       366
Attributes:      10
                 action
                 difficult
                 std_ability
                 answer_incorrect
                 answer_partial
                 hints_info
                 hints_partial
                 time
                 pknow
                 class
Test mode:       evaluate on training data

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

action = comfirm_answer_partial
|   time = low
|   |   answer_partial <= 0: no_report (39.0)
|   |   answer_partial > 0: increase (4.0)
|   time = medium: increase (18.0/2.0)
|   time = high: increase (20.0/4.0)
action = comfirm_answer_incorrect
|   answer_incorrect <= 0
|   |   hints_info <= 0: no_report (26.0/6.0)
|   |   hints_info > 0
|   |   |   answer_partial <= 0: no_report (10.0/1.0)
|   |   |   answer_partial > 0
|   |   |   |   time = low: decrease (3.0/1.0)
|   |   |   |   time = medium: no_report (4.0/1.0)
|   |   |   |   time = high: no_report (5.0/2.0)
|   answer_incorrect > 0
|   |   time = low: no_report (8.0/3.0)
|   |   time = medium: increase (10.0)
|   |   time = high: increase (9.0)
action = hint_info
|   answer_incorrect <= 1
|   |   need4help = yes
|   |   |   hints_info <= 0: no_report (11.0/2.0)
```

```
|   |   |   hints_info > 0
|   |   |   |   answer_partial <= 0: no_report (3.0/1.0)
|   |   |   |   answer_partial > 0: increase (4.0/1.0)
|   |   need4help = no: no_report (30.0)


|   |   need4help = yes
|   |   |   answer_incorrect <= 0
|   |   |   |   hints_info <= 0: no_report (5.0/2.0)
|   |   |   |   hints_info > 0
|   |   |   |   |   answer_partial <= 0: no_report (3.0/1.0)
|   |   |   |   |   answer_partial > 0: increase (4.0/1.0)
|   |   |   answer_incorrect > 0: no_report (6.0)
|   |   need4help = no: no_report (30.0)
|   answer_incorrect > 1
|   |   difficult <= 1: decrease (14.0)
|   |   difficult > 1
|   |   |   difficult <= 2
|   |   |   |   time = low: decrease (4.0)
|   |   |   |   time = medium: no_report (24.0)
|   |   |   |   time = high: increase (2.0)
|   |   |   difficult > 2: no_report (10.0)
action = hints_partial
|   hints_partial <= 1: no_report (16.0)
|   hints_partial > 1: increase (8.0)
action = comfirm_answer_correct
|   hints_info <= 1: increase (27.0)
|   hints_info > 1: no_report (27.0)
action = prompt_exit_theory_example: no_report (19.0/1.0)
action = prompt_exit_other
|   benefInter = yes: no_report (7.0/1.0)
|   benefInter = no: decrease (7.0/1.0)
action = prompt_stay
|   benefInter = yes: no_report (11.0/3.0)
|   benefInter = no: increase (8.0)

Number of Leaves  :     32

Size of the tree :      51


Time taken to build model: 0.02 seconds

=== Evaluation on training set ===
=== Summary ===

Correctly Classified Instances         340               92.8962 %
Incorrectly Classified Instances        26                7.1038 %
Kappa statistic                          0.868
Mean absolute error                      0.0744
```

```
Root mean squared error                  0.1893
Relative absolute error                 20.219  %
Root relative squared error             44.178  %
Total Number of Instances              366

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.94      0.024     0.948     0.94      0.944     increase
  0.962     0.118     0.919     0.962     0.94      no_report
  0.703     0.006     0.929     0.703     0.8       decrease

=== Confusion Matrix ===

   a   b   c   <-- classified as
 109   7   0 |   a = increase
   6 205   2 |   b = no_report
   0  11  26 |   c = decrease
```

D.3 Decision Trees - Tutor Perspective

D.3.1 Confidence

```
Scheme:       weka.classifiers.trees.J48
Instances:    110
Attributes:   26
   Delta2
              action
              Difficulty
              statement/no_know
              accepthelp
              clarification
              comprehension
              confirm/answer
              confirm/planning
              planning/student/continue
              disjuctive
              expectational
              instrumental
              planning/positive
              planning/negative
              query
              reflection
              request/help
              selfcorrection
              statement/affect
```

```
               answer_correct
               answer_partial
               answer_incorrect
               aside
               acknowledgment
               statement/know

Test mode:     10-fold cross-validation

=== Classifier model (full training set) ===

action = answer_Correct: increase (36.0/6.0)
action = confirm/answer
|   Difficulty = Medium: decreasing (8.0/1.0)
|   Difficulty = Low
|   |   answer_partial <= 0: init_Medium (6.0/2.0)
|   |   answer_partial > 0: decreasing (3.0/1.0)
|   Difficulty = High: decreasing (2.0/1.0)
action = statement/no_know
|   statement/no_know <= 0
|   |   Difficulty = Medium: init_Medium (3.0)
|   |   Difficulty = Low: init_Low (3.0/1.0)
|   statement/no_know > 0
|   |   answer_correct <= 0: decreasing (4.0/1.0)
|   |   answer_correct > 0: extrm_decreasing (2.0)
action = answer_Partial
|   answer_partial <= 0
|   |   Difficulty = Medium: decreasing (1.0)
|   |   Difficulty = Low: increase (4.0/2.0)
|   |   Difficulty = High: increase (3.0)
|   answer_partial > 0: decreasing (4.0/1.0)
action = query/factoid: decreasing (1.0)
action = acknowledgment: decreasing (2.0/1.0)
action = planning/negative: decreasing (1.0)
action = answer_Incorrect
|   query <= 0
|   |   Difficulty = Medium: increase (4.0)
|   |   Difficulty = Low: extrm_decreasing (4.0/2.0)
|   |   Difficulty = High: increase (2.0/1.0)
|   query > 0: decreasing (2.0/1.0)
action = request/help: decreasing (1.0)
action = reflection: increase (2.0)
action = planning/positive: increase (2.0)
action = selfcorrection: increase (3.0)
action = confirm/planning: increase (1.0)
action = statement/know: init_High (5.0)

=== Stratified cross-validation - Summary ===

Correctly Classified Instances          61.8182 %
```

```
Incorrectly Classified Instances        38.1818 %
Kappa statistic                          0.4393


=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.346     0.095      0.529     0.346     0.419    decreasing
  0         0.058      0         0         0        extreme_decreasing
  0.882     0.305      0.714     0.882     0.789    increase
  0.636     0          1         0.636     0.778    init_High
  0         0.009      0         0         0        init_Low
  0.385     0.113      0.313     0.385     0.345    init_Medium

=== Confusion Matrix ===

  a  b  c  d  e  f   <-- classified as
  9  3  9  0  0  5 |  a = decreasing
  3  0  1  0  1  2 |  b = extreme_decreasing
  2  0 45  0  0  4 |  c = increase
  0  0  4  7  0  0 |  d = init_High
  0  2  0  0  0  0 |  e = init_Low
  3  1  4  0  0  5 |  f = init_Medium
```

D.3.2 Effort

```
=== Run information ===

Scheme:       weka.classifiers.trees.J48 -U -M 2
Relation:     effort_summary_statics
Instances:    365

Attributes:   effort_change, time, need4hint, diff, std_ability,
              hist_vector {answer_incorrect, answer_partial}

Test mode:    10-fold cross-validation

=== Classifier model (full training set) ===

J48 unpruned tree
------------------

answer_correct <= 0
|   statement/no_know <= 0
|   |   answer_incorrect <= 0
|   |   |   action = accepthelp: static (0.0)
|   |   |   action = acknowldgment: static (0.0)
|   |   |   action = acknowledgment: static (1.0)
```

```
|   |   |   action = acknowledgment/answer: static (1.0)
|   |   |   action = answer_Correct
|   |   |   |   Difficulty = Irrelevant: init_High (3.0/1.0)
|   |   |   |   Difficulty = Medium
|   |   |   |   |   query <= 0: init_Medium (5.0/3.0)
|   |   |   |   |   query > 0: static (2.0)
|   |   |   |   Difficulty = High: static (0.0)
|   |   |   |   Difficulty = Low: init_High (6.0/3.0)
|   |   |   |   Difficulty = VeryHigh: static (1.0)
|   |   |   |   Difficulty = MediumHigh: static (0.0)
|   |   |   action = answer_Partial
|   |   |   |   Difficulty = Irrelevant: static (0.0)
|   |   |   |   Difficulty = Medium: init_High (1.0)
|   |   |   |   Difficulty = High: init_High (1.0)
|   |   |   |   Difficulty = Low: static (2.0)
|   |   |   |   Difficulty = VeryHigh: static (0.0)
|   |   |   |   Difficulty = MediumHigh: static (2.0)
|   |   |   action = answer_Incorrect
|   |   |   |   Difficulty = Irrelevant: static (2.0)
|   |   |   |   Difficulty = Medium: init_High (1.0)
|   |   |   |   Difficulty = High: static (2.0/1.0)
|   |   |   |   Difficulty = Low: init_High (1.0)
|   |   |   |   Difficulty = VeryHigh: static (0.0)
|   |   |   |   Difficulty = MediumHigh: static (0.0)
|   |   |   action = answer_Irrelevant: init_High (1.0)
|   |   |   action = clarification: static (0.0)
|   |   |   action = aknowldgement: static (0.0)
|   |   |   action = confirm/answer: static (4.0/2.0)
|   |   |   action = confirm/planning: static (1.0)
|   |   |   action = expectational: static (1.0)
|   |   |   action = planning/negative: static (0.0)
|   |   |   action = planning/positive: static (0.0)
|   |   |   action = planning/student: static (0.0)
|   |   |   action = query/factoid: static (0.0)
|   |   |   action = query/process/how: static (2.0/1.0)
|   |   |   action = query/process/what: static (0.0)
|   |   |   action = reflection: static (0.0)
|   |   |   action = request/help: static (0.0)
|   |   |   action = selfcorrection: static (0.0)
|   |   |   action = statement/affect: static (0.0)
|   |   |   action = statement/no_know: static (3.0/2.0)
|   |   |   action = teaching: static (0.0)
|   |   |   action = tidy: static (0.0)
|   |   answer_incorrect > 0: static (15.0/2.0)
|   statement/no_know > 0
|   |   action = accepthelp: static (0.0)
|   |   action = acknowldgment: static (0.0)
|   |   action = acknowledgment: static (0.0)
|   |   action = acknowledgment/answer: static (1.0)
|   |   action = answer_Correct: increase (3.0/1.0)
```

```
|   |   action = answer_Partial: static (5.0/1.0)
|   |   action = answer_Incorrect: static (3.0)
|   |   action = answer_Irrelevant: static (0.0)
|   |   action = clarification: static (0.0)
|   |   action = aknowldgement: static (0.0)
|   |   action = confirm/answer: static (5.0/1.0)
|   |   action = confirm/planning: static (1.0)
|   |   action = expectational: static (0.0)
|   |   action = planning/negative: static (0.0)
|   |   action = planning/positive: static (0.0)
|   |   action = planning/student: static (0.0)
|   |   action = query/factoid: static (0.0)
|   |   action = query/process/how: static (0.0)
|   |   action = query/process/what: static (1.0)
|   |   action = reflection: static (0.0)
|   |   action = request/help: static (0.0)
|   |   action = selfcorrection: static (0.0)
|   |   action = statement/affect: static (1.0)
|   |   action = statement/no_know: static (2.0/1.0)
|   |   action = teaching: static (0.0)
|   |   action = tidy: static (0.0)
answer_correct > 0
|   action = accepthelp: static (1.0)
|   action = acknowldgment: static (0.0)
|   action = acknowledgment: static (2.0)
|   action = acknowledgment/answer: static (5.0)
|   action = answer_Correct
|   |   acknowledgment <= 2
|   |   |   answer_correct <= 7: static (94.0/3.0)
|   |   |   answer_correct > 7
|   |   |   |   statement/affect <= 0: increase (2.0)
|   |   |   |   statement/affect > 0: static (4.0)
|   |   acknowledgment > 2: static (3.0/1.0)
|   action = answer_Partial: static (31.0)
|   action = answer_Incorrect: static (31.0/1.0)
|   action = answer_Irrelevant: static (13.0)
|   action = clarification: static (3.0)
|   action = aknowldgement: static (2.0)
|   action = confirm/answer: static (32.0/2.0)
|   action = confirm/planning: static (7.0/1.0)
|   action = expectational: static (1.0)
|   action = planning/negative: static (1.0)
|   action = planning/positive: static (3.0)
|   action = planning/student: static (3.0)
|   action = query/factoid: static (2.0/1.0)
|   action = query/process/how: static (1.0)
|   action = query/process/what: static (0.0)
|   action = reflection: static (3.0)
|   action = request/help: static (3.0/1.0)
|   action = selfcorrection: static (4.0)
```

```
|   action = statement/affect: static (4.0)
|   action = statement/no_know
|   |   confirm/planning <= 0: static (15.0/2.0)
|   |   confirm/planning > 0: decreasing (3.0/1.0)
|   action = teaching: static (3.0)
|   action = tidy: static (9.0)

Number of Leaves  :     99

Size of the tree :      113


Time taken to build model: 0.05 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances         304               83.2877 %
Incorrectly Classified Instances        61               16.7123 %
Kappa statistic                          0.1528
Total Number of Instances              365

=== Detailed Accuracy By Class ===

TP Rate   FP Rate   Precision   Recall  F-Measure   Class
  0.94      0.739      0.898     0.94      0.919    static
  0         0.014      0         0         0        decreasing
  0         0.006      0         0         0        increase
  0.267     0.046      0.2       0.267     0.229    init_High
  0         0          0         0         0        init_Low
  0         0.011      0         0         0        init_Medium

=== Confusion Matrix ===

   a   b   c   d   e   f   <-- classified as
 300   3   1  13   0   2 |   a = static
   9   0   1   0   0   0 |   b = decreasing
  12   2   0   0   0   0 |   c = increase
  10   0   0   4   0   1 |   d = init_High
   0   0   0   1   0   1 |   e = init_Low
   3   0   0   2   0   0 |   f = init_Medium
```

Appendix E

Recording, Replaying and Visualising Students' Behaviours in ILE

This Appendix describes briefly how students actions were recorded and replayed. The following are adapted from Mavrikis (2005).

E.1 Recording students' actions

There are several ways to record students' actions when they are working in a predefined environment or a lab where an experiment or study is set up. On the other hand, it is very difficult to set up and use equipment (such as cameras, tape recorders or even screen capture software) let alone have an observer to study what students are doing when they are interacting in their time and place. Of course, one can analyse the log files of the web server that provide information on the pages and the time they were accessed but, especially when the environment involves client-side interactions, the data logged can be quite limited for the research purposes in the field of AIEd and ITS.

In order to log and replay the students interaction with the web-based system an agent that can be added to an ILE was developed. The agent records every student action including mouse movements, mouse and keyboard clicks etc. A piece of JavaScript code adapts the elements (buttons, divs, input boxes etc.) of every (X)HTML page that the system delivers adding event handlers to all of them so that

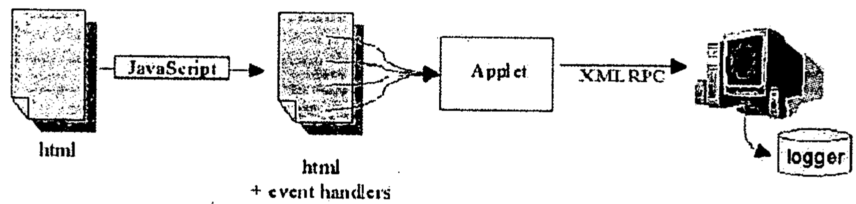


Figure E.1: The architecture of the agent which logs students' interactions

they can send messages to a hidden Java applet[1]. The applet in turn sends them, every so often (or just before the student logs out), to a server where they are logged (see Figure E.1). The data logged serve several purposes and provide the opportunity not only to replay accurately the student's interaction with the system but also permit data-mining of information that would not be accessible otherwise (e.g., cancelled submissions or clicks, or deleted typing that can be interpreted as hesitation or self-correction).

It is worth noting, for the benefit of future research, that there are several technical problems with the aforementioned solution as it is constrained by bandwidth availability, several security settings that have to be met, as well as javascript-applet communication which is neither reliable, nor standardised. This resulted, in few cases, in very noisy data that had to be discarded. Most of the constraints were possible to overcome during this thesis due to fortunate fact that most students work with the system under the university network which allows for several security constraints to be dropped in terms of logging. Even when they are at their own place and time, most live in university halls of residence that still belong to the same network. The last few years the advancements in client-server communication have made possible the implementation of similar solutions using asynchronous javascript (see AJAX - [2])

[1]Note that, although the agent resembles "spyware" it was not designed to be such. The students are aware of the data collection and have to agree (or disagree) explicitly to their behaviour being recorded.

[2]http://developer.mozilla.org/en/docs/AJAX:Getting_Started

E.2 Replaying and analysing the log files

As already mentioned, the data collected involve every student action from page selection to page abandonment and from mouse movements and clicks to key presses. Being able to replay students' full interaction opens several possibilities of knowledge elicitiation from experts, tutors and the students themselves. In addition developing the agent for the logging is a step towards any agent that needs to be added to the ILE to inform the student model about the actions the student takes.

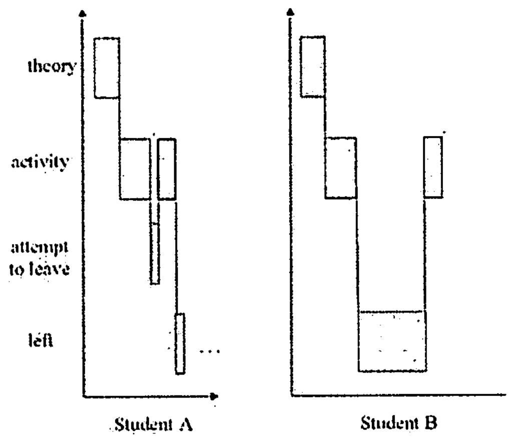


Figure E.2: Navigation actions

Developing the right tools and being able to replay and visually represent the data was a precursor to analysing them in a coherent way. The most difficult part, however, was to identify the best level of abstraction for the various interactions and determine the most interesting ones to replay. An ad-hoc parser was developed,that depending on the level of the desired analysis, processes the log files and has the ability to replay the interaction[3]. In addition, the parser builds an output of the interactions of interest in a structure that can be read by Matlab to produce graphs that present a students' interaction in a visual way and make it easier to choose particular places to replay.

[3]The Java Robot package is used for the purposes of the replay. Particular atttention was paid to record the exact time it took pages to load to the students' browser in order to be able to synchronise with the timestamped actions. Of course this is also subject to noise and other network problems leading to further data loss. The problems can be rectified by adding some process in between that verifies the packets transmitted between server and client.

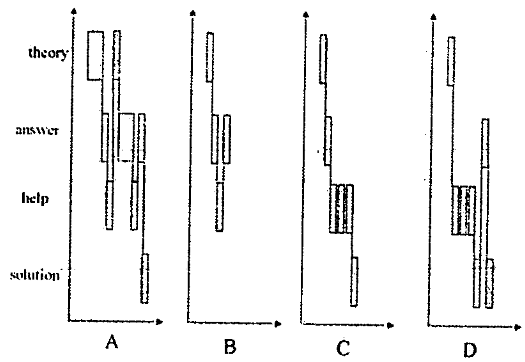


Figure E.3: Local actions

As an example, Figure E.2 illustrates two users with different behaviour. Student A attempted to leave but under the system's guidance (see Section 5.4.3.2) remained in the page to complete the activity. Student B left, visited other pages and then returned to complete it. Similarly, local actions may also have an effect on diagnosis. For example, the interactions of student A, C and D (Figure E.2) are quite different despite the fact that they all ask for the solution at the end of the activity. Finally, having the data in a machine analysable form makes it easier to search for patterns of interactions such as holding down a button and then releasing it after a while (which shows signs of hesitation) or cancelling the decision to click the submit button. These would otherwise be impossible.