Activity Theory based evaluation of DMSS–R

Lars Larsson

June 13, 2008
Master's Thesis in Computing Science, 30 credits
Supervisor at CS-UmU: Helena Lindgren
Examiner: Per Lindström

Umeå University
Department of Computing Science
SE-901 87 UMEÅ
SWEDEN

Abstract

The Dementia Management and Support System Research and Evaluation version (DMSS-R), developed at the Department of Computing Science at the University of Umeå was evaluated using a group of medical students, an intern and an expert. The goal of the study was to, from a non-expert perspective, study how the system may be improved to offer better support for medical reasoning and as a learning tool, in addition to how the user interface may be improved for increased intuitiveness. Data was gathered in the form of observations of sessions with real and fabricated patient cases. These were supplemented with interviews and by asking the test subjects to “think aloud”. The data from the qualitative study was analysed using Activity Theory, supplemented by heuristic evaluation and cognitive load theory. Breakpoints were identified, and suggestions for remedying these are suggested. Due to low number of test subjects, the results of the study should be seen as indicative only. The main result of the study is that DMSS-R must offer explanatory feedback to promote learning and thus improved medical reasoning support for non-experts, in addition to several user interface changes.

Additionally, the attitudes toward the system were studied, and the results show that in spite of not being convinced by the system-suggested diagnosis, the test subjects are overall very positive toward using the system and think that it may provide valuable support in the future.

Contents

List of Figures

1 Introduction

Dementia care in Sweden is in need of improvement. A study by Ólafsdóttir in 2001 (as referenced in [24]) showed that a mere 25 per cent of patients with dementia were detected by their general practitioner during their first appointment. The majority of patients with dementia in the study (66 per cent) had been misdiagnosed with some psychiatric disease or problem, such as sleep disorder or anxiety. To remedy this situation, interest groups and experts have established a goal for the improvement of Swedish dementia care — the goal being that 90 per cent of dementia patients, rather than the current 25 per cent, should be properly diagnosed at their first meeting with their general practitioner. Clearly, this goal is ambitious. But is it unattainable? Using the right tools, and proper training, it should be possible to overcome the difficulties of today and improve the dementia care of tomorrow.

This Master of Science thesis evaluates a computer-based decision support system called DMSS-R, short for Dementia Management and Support System Research and Evaluation version. DMSS-R is being actively developed at the Department of Computing Science at the University of Umeå. The goal of the system is to provide one of the tools needed to bridge the gap between dementia care in its current and future states.

After a short introduction to the medical domain in Section 1.1, describing dementia diseases and their effects in various contexts in brevity, we focus attention on some of the sources of problems and solutions in modern dementia care. Decision support systems are inherently dependent upon cognitive science, if not to produce valid logic, then at least as they serve as cognitive tools for physicians who use them. Section 1.2 is concerned with knowledge building and representation, and hypotheses generation. The section is followed by a discussion on clinical guidelines as a means of raising the standards in clinical decision making in Section 1.3. DMSS-R is at heart a large set of guidelines encoded into formal logic, thus making it important to compare this computer-based tool to ones already in use — and to learn from studies performed on other guideline representations.

1.1 Brief introduction to the medical domain

As we grow older, our intellectual functioning decreases. This is intuitively understood, and has also been shown in studies by, among others, Thorndyke 1928 and Wechsler 1958 (as referenced in [3]). This means that intelligence, or cognitive performance, decreases with age starting at around age 40 by a slow decline and increasing rapidly from age 60 until death [3]. This decrease is natural and not symptomatic of a disease. However, dementia is a group of diseases that cause an unnatural decrease of cognitive functions and other brain functions. Some of these diseases are very hard to differentiate, thus making it hard to engage the proper treatment in a timely manner [11].

As a cognitive disease, dementia includes brain damages of various degrees. These damages are usually cognitive disorders, neurological dysfunctions, and are often followed by psychological changes [24]. The cognitive disorders include, but are not limited to, loss of memory — a symptom usually noticed quickly by immediate family. Other disorders include higher level reasoning skills, abstract thinking, and communication skills. Since these skills are vital to performing well in activities of daily life (ADL), dementia patients become more dependent on others as the disease progresses. This usually places a large burden on the family of the patient, and a financial one on society, as the need for care increases.

In addition to a decreasing ability to perform ADL, people suffering from dementia experience behavioural and psychological symptoms (BPSD — behavioural and psychological symptoms in dementia). These include hallucinations, depression, sleep difficulties, anxiety, becoming highly emotional and misconceptions. Common misconceptions are that others are stealing, the life partner is cheating, and so forth. Coupled with the higher dependency on others, and the realization that "something is wrong", many patients with dementia develop anger problems and can become quite aggressive. In fact, increased aggression is one of the most common reasons why people (are forced to) seek professional help. Other problems, such as depression and anxiety, often go untreated for longer, under the misconception that they are simply parts of growing older.

Determining that a person suffers from dementia is, as the previously referenced study by Ólafsdóttir indicates, not a simple task. Some of the difficulty is due to the natural decline in cognitive functions that makes it hard to correctly determine that a state of dementia is present [51]. In addition, it may be due to low levels of experience in the general practitioners — a general practitioner meets about 1-4 new dementia patients per year [24]. These low numbers, combined with new guidelines and treatment strategies being developed constantly and a high turn-over of medical staff, increase the difficulty of the situation.

Working toward increasing efficiency in the care of dementia patients has many good effects. Not only for the patients, but society at large. A 2005 study by Wimo et al. states that approximately 50 billion SEK are spent every year on patients with dementia [52], the vast majority of that being costs associated with care — not medication. Medication is fairly cheap, but effective, for some types of dementia [51]. By slowing the detrimental aspects of the disease down, the patient can gain from months to years of relatively unaffected living. Kelly et al. showed in a study from 1997 (as referenced in [24]) that treating Alzheimer's disease, one of the most common types of dementia, with drugs could delay institutionalization of patients with up to nine months, reducing costs by 17–30 per cent.

Seen in light of the problems presented above, the need for better routines, processes, and supporting tools is apparent. In order to develop such tools, and study them for evaluation purposes, we need to consider the cognitive processes of the physicians, rather than those of the patients. With better understanding comes possibilities of supporting these processes better.

1.2 Knowledge building and representation

The medical domain is divided into many different areas and physicians working in medicine are divided by area of expertise. Some choose to become experts within a single area, such as geriatrics, whereas others choose the broad area of general practice. As a patient, the first and foremost contact with medical personnel is usually with a general practitioner. As stated in the previous section, a Swedish general practitioner may meet between 1 and 4 new dementia patients per year [24]. A mere quarter of these get the proper diagnosis, which leads to questions concerning how to raise the level of expertise. In order to answer these, we direct attention to how knowledge is built and represented, and how hypotheses are generated. We shall see that there is a difference between experts and non-experts in these areas. If this gap could be closed, for instance by using some tool such as DMSS-R, the difference would be less significant and thus more patients would benefit from expert-level advice. Before looking at some differences between novices and experts, where novices is understood as including people who may be experts in some field but not in the one in question, we turn our attention to results in decision-making research that are common to all humans.

Medical decision research has focused on two objectives: understanding how decisions are made in "real-world" settings by physicians, other health care personnel, and patients, and creating ways of supporting the decision-making process. The latter includes the creation of guidelines (the topic of Section 1.3) and other types of decision-support systems. There have been different approaches to the research of decision making over the last few years, each with its own merit. The "traditional" decision research is concerned with topics such as rationality, biases, utilities and heuristics. The traditional view is discussed briefly below. However, Patel et al. reason that traditional decision research is neither able to adequately guide the development and implementation of decision support systems, nor can it be used to improve the practice of evidence-based medicine [41].

Reducing decision making to its core, a decision involves a choice of actions and options, beliefs about objective states and events in the world (including outcomes and means to achieve them) and some sort of desire or motive to move in some direction, as the driving force to make a decision in the first place [13]. By that definition, a good decision is one that involves making a good choice of action or option, that reaches the desired goal state and is in accordance with the beliefs about the state of the world.

Focusing on the choice of action or option, how can such a choice be made? When is it good? Rational choice theories attempt to answer these questions. If we can somehow assign scores to different choices, choosing the one which maximizes our score would be the most rational. Normative theories, of which rational choice theory is one, are based on two main types of models. The first of these are those of expected utility (EU) and subjective expected utility (SEU) [41]. Both use the idea that one should maximize personal gain, calculated as the ration of chance taken by the amount of payoff. The second common normative model is based on conditional probability, expressed in subjectivist, personalist, or Bayesian perspective [41]. One of the strengths of these models is that they provide a mathematical foundation, and can thus be tested in laboratory settings. Using logic, it is possible to know beyond the shadow of a doubt what the correct answer to a question is — thus rendering all other answers provably incorrect. While these models have problems that will be discussed later, they have influenced the subsequent theories and models in decision research and are thus still important.

1.2.1 Deviations from rational choice

Biases and heuristics have been shown to be important to decision-making (Chapman and Elstein, as referenced in [41]). Using the normative approach, one can define biases as systematic deviations from the normative standards. It is important to study biases, as they both give insight into cognitive processes underlying decision-making and may provide valuable suggestions of areas in which improvement in decision-making can be made. The study of biases as sources of error in human reasoning has amassed a large body of research. Humans are simply not perfectly rational beings. Amos Tversky and Daniel Kahneman studied, among other things, human ability to assess probabilities of uncertain events or values of uncertain quantities. They discovered heuristics (or biases) that they claimed were generally useful, but could lead to severe and systematic errors [47]. Using a simple example where, based on a story and known percentages for the likelihood of relevant facts, test subjects were supposed to estimate the likelihood of some statements, Tversky and Kahneman showed that probabilities of events are estimated from a population according to the representativeness of the sample. This leads to faulty reasoning, which is especially dangerous in the medical domain where probabilities are used often to describe likelihood of certain symptoms appearing for a given disease. As a telling example of this, consider the following (Eddy 1982, as referenced in [41]):

Estimate the probability of a woman having breast cancer, given that her mammogram results are positive and the following information:

– the prior probability that a patient has breast cancer is 1 per cent;
– if a patient has breast cancer, the probability of a correct diagnosis (positive mammogram) is 80 per cent; and
– if the patient has a benign lesion (no breast cancer), the probability that the mammogram is positive is 9.6 per cent (false positives).

Eddy found that 95 of 100 physicians estimated the probability of breast cancer after a positive mammogram at 75 per cent. Mathematically, one can easily apply Bayes' rule and show that the correct answer is a mere 8 per cent probability. Other studies showing the same result have also been conducted.

In addition to our poor ability to make mathematically sound decisions based on statistics, common biases have been documented. *Availability bias* is the tendency to assess probabilities or likely causes of events by how readily available the event is in memory. If a patient exhibits very similar symptoms to another patient, whom the physician remembers in a very vivid way and was successfully treated with some remedy, the physician is likely to assume that the same treatment works for the new patient. This is reminiscent of mental models [44], and is a form of heuristics that is commonly used by experts. However, it may lead to fatal decisions, such as giving the wrong medication to patients with a different type of dementia, in effect shortening their life dramatically [51].

Researchers have found many types of heuristics and biases that are used irrationally by humans (notably Chapman and Elstein, as referenced in [41]). While some are intuitively obvious (such as overconfidence and hindsight), two that are particularly problematic in medical settings are *confirmation bias* and the *framing effect.* Confirmation bias can be intuitively understood as the force that keeps us from deviating from our hypothesis — once we have chosen a hypothesis, we tend to view all new evidence that supports it as more important than evidence that suggests alternatives. This can lead to incorrect diagnosis.

The framing effect relates to the preference of some outcomes over others. Given two possible treatments, the way in which they are expressed (or framed) can highly influence which will be chosen. In suggesting a treatment, expressing the effects in terms of survival rate or mortality rate can have a large impact [31]. While survival and mortality rate are merely two sides of the same coin, positive framing leads to risk-averse choices, and vice versa.

The lesson to be learned from these deviations from rationality are that a decision support system should be developed with these in mind, as the final judgement will always be made by a skilled physician. Biases such as the confirmation bias or the framing effect must be carefully avoided from the point of view of the system, lest it be biased itself. As shown by Eddy, statistical data may also be problematic when interpreted by humans. Gigerenzer (2000, as referenced in [41]) showed that, using the same experiment that Eddy used, but with the percentages replaced by phrases such as "ten out of every 1000", the amount of correct responses went from a paltry 5 per cent to 46. While the experiment has been criticized, it does seem to indicate that language plays a large part in our understanding of even something "pure" as mathematics.

1.2.2 Hypothesis generation and information storage

Patel et al. argue that, in order to understand how humans reason, we should direct studies toward problem-solving research rather than decision-making research [41]. The arguments laid forth will not be presented here, but one of the main differences between the two approaches is that problem-solving research views the decisions of experts as the "gold standard", whereas decision-making research (as discussed in the previous section) considers anything that is not mathematically proved to be the single rational choice to be an error in human reasoning. Evidence-based medicine is an attempt of solving the problems of contemporary medical care, by applying the scientific method to certain aspects of the domain. It is, however, incomplete. Experts, then, are seen as the source of correct knowledge. But, as shown by Olsén, experts are not perfectly reliable. Domain experts are not automatically guaranteed to reach the same conclusions [38]. Still, given that humans have been shown to not be rational beings, it seems more valuable to aspire to reaching the skill level of expert (and as such, make the "right" choice in *almost* all cases, even though differences between experts may exist) rather than attempting to be something we are not.

In problem solving research, the process of solving a problem is seen as operations performed by an actor, the problem solver, in order to work toward a goal state in a space of possible operations. The emphasis is placed not only on reaching the goal state itself, but also on the path toward it. Solving a problem is an evolving process, and cannot be understood solely in mathematical terms.

A cornerstone in research employing the problem-solving approach to study clinical decision-making and practice is that physicians, experts and novices alike, generate hypotheses and use data from test results to refine the set of possible hypotheses to find the correct one [41]. This is precisely the problem with what decision research calls the confirmation bias (see previous section). Does this imply that physicians employing this technique are wrong and their work flawed? Several studies have shown that this is not the case [41].

To understand why, we turn to artificial intelligence (AI). As anyone who has played chess knows, the number of possible operations even in a highly limited and constrained situation as those on the chess board is staggering. Without some useful plan, some heuristic that guides the reasoning process, it would simply be impossible to begin making even the first move. In AI, this technique of applying some heuristic to reduce the amount of work that needs to be done is known as *pruning*, in which unfavourable paths (sequences of decisions) disregarded as early as possible. The same process seems to be present in human reasoning. Pattern recognition, a skill that humans possess and is non-trivial to teach computers, is another aspect of the same phenomenon. Experts have been shown to have very highly developed pattern recognition, or pruning, skills and focus their attention on the task at hand quickly. Klein et al. (1995, as referenced in [28]) showed, via think-aloud protocol, that skilled chess players generated useful moves as the first ones they considered. The search was thus not random through the large space of possible moves, but guided by expertise on part of the players. The same has been found by Patel et al. in the medical domain [41].

One of the main differences between the expert and the non-expert is that experts are more skilled in generating the correct hypothesis in the set of hypotheses, and that once it has been generated it will be confirmed with more accuracy than by non-experts [41]. Even if the non-expert generates the correct hypothesis, not only will the set of incorrect hypotheses be larger, the final decision takes longer as the non-expert is less skilled in eliminating the incorrect ones.

In other research, it has been shown that non-experts are more inclined to suggesting treatment on insufficient grounds than experts, who instead preferred to stabilize the patient and gather more data [41]. This result, combined with the previously stated that non-experts generate more (incorrect) hypotheses clearly show a problem that can be alleviated using cognitive tools such as decision support systems.

1.3 Clinical guidelines and their use in health care

Decisions in the medical domain are intrinsically hard, as they, in a very real sense, are matters of life and death. The hard decisions are made even harder by the constant evolution of medical knowledge and science. To alleviate the task of making these decisions, and to ensure that efficient and cost-reducing practices are employed, medical guidelines have been developed. A modern form of these guidelines are decision support systems, whose goal it is to present the guidelines in an intuitive and easy-to-use way.

Patel et al. have showed that experts and non-experts (as defined as the complement of experts, including general practitioners and specialists working outside their field of expertise) use guideline support differently [40]. Since DMSS-R is intended for a broad spectrum of users, we shall therefore study these differences, as they are of great importance to the kind of support DMSS-R offers. Patel et al. cover five differences in comprehension, problem-solving and decision-making between experts and non-experts [40]. These are:

1. Differences in patterns of reasoning. Experts employ a data-driven method of reasoning, whereas non-experts use a hypothesis-driven approach. Experts, therefore, study data and create a hypothesis as a result. Non-experts generate a hypothesis (or a set of hypotheses) prematurely and try to find support for it, and interpretation models, in the data presented to them.

2. Knowledge base organization differences. Experts have a highly organised knowledge base, and while non-experts may also have a large knowledge base, theirs is not as highly organised. This results in generation of hypotheses that are not relevant to the problem at hand. In conjunction with the previously stated difference of non-experts reasoning in a hypothesis-driven manner, this creates problems of inefficiency.

3. Differences in the way that errors are made. Non-experts are guilty of making errors of omission, largely due to inability to separate useful from irrelevant information. Experts, on the other hand, make the same kind of errors due to over-confidence or as a result from skipping steps in problem solving.

4. Different approach to clinical problems. Experts generate a small set of hypotheses, perhaps only a single one, using reasoning on a high level of abstraction and can quickly narrow this set down to the most accurate. In cases where time is a factor, experts also employ a set of rules of thumb — something that non-experts cannot do, due to a lack of experience.

5. Differences in flexibility and experience. Non-experts often lack experience and are thus more dependent on scientific evidence, using it more often than experience-nurtured intuition. This makes them less able to deal with unique or novel cases. The experts, however, may use experience instead of (or in conjunction with) scientific evidence to make a correct and informed decision.

These differences between the experts and non-experts come with implications for how guidelines may be used in practice, and even for how they are written and presented. Due to the fact that guidelines are generally written by experts, they may be written in a way that is inaccessible to non-experts [40]. Patel et al. also showed in a study that the choice of representation is important when it comes to supporting medical decisions via guidelines, indicating that algorithmic representations were superior to textual ones [40]. The superiority was manifested in several ways, and most importantly, the number of irrelevant tests ordered by the physicians was greatly decreased, directing the medical investigation toward the goal and thus making it more efficient.

Patel et al. pose two hypotheses that are troubling: experts may avoid guideline usage because they find that the guidelines do not add anything to their practice, and non-experts use them inaccurately (if at all) [40]. If these hypotheses are indeed true, there are clearly not only technical difficulties that need to be overcome in the development of a computer decision support system. Overcoming these difficulties could, however, lead to great gains — the system could function not only as a tool for learning for non-experts, but as a reminder for experts who are used to rely on their expertise, which, as times go by, might become outdated and incorrect. It has been shown that experts tend to use their expertise to *satisfice* rather than maximise [40] and improvise [8] using their large knowledge bases. These strategies may work well in the majority of cases, but do not guarantee that all patients receive equal treatment. In a universal health care country such as Sweden, this is not only important on a personal level, but it is also a political issue.

1.4 Thesis topic

Patel et al. in the discussion of medical guidelines [40, page 164], state that:

> Guideline delivery forms that are dynamically generated and adaptable to the expertise level of the user may be more effective [than previous forms of representation]. Neither text-based nor diagrammatic forms alone could serve this adaptive purpose. For clinical practice guidelines to be successfully used, either as decision-making or as educational tools, they need to be included in decision support systems [. . .] that help focus the non-expert physician on relevant information or remind the expert physician of important steps.

This shows the importance of decision support systems, and that there is a perceived need for them in the medical domain. The topic of this thesis is to evaluate how well suited DMSS-R is as a tool for use by non-experts, with focus on the following questions:

1. In what ways can DMSS-R be improved to better support the medical reasoning and investigation processes?
2. How can the graphical user interface of DMSS-R be improved to increase intuitiveness and ease-of-use?
3. How may DMSS-R be improved as a learning aid?

A recent study by Lindgren shows that the system is lacking in certain aspects as a support for medical reasoning processes [26]. The study gave rise to the topic of this thesis. Similarly, another study by Lindgren highlights problems that may arise due to cultural context, as DMSS-R is being developed for both Swedish and Asian markets [25]. Other studies have focused more on the underlying system logic, such as the one by Lindgren and Eklund [27] and Eklund et al. [9]. One thing these previous studies all have in common, in spite of their varied topics, is that none of them have focused on usability concerns. This thesis aims to help the developers of the system by identifying the major sources of interface and conceptual breakdowns for non-expert users (note that the referenced studies have been conducted using medical experts who were also expert users), which is of particular interest due to a planned study for the autum of 2008 where DMSS-R is being deployed for use by general practitioners. The results of this thesis are intended to improve the quality of the system and minimise obstacles that could negatively influence the upcoming study.

As suggested by Patel et al., physicians at different levels of expertise may require different kinds of support, presented in different ways. The questions are all answered from the *point of view of the non-expert.* As stated in the beginning of the chapter, non-experts represented by the general practitioners have a low success rate of determining the correct type of dementia in patients. Therefore, supporting the decisions of non-experts is seen as one of the areas in which improvement is most required.

It will not be possible to answer the questions fully, given their open-ended nature. In particular, question number 3 is deemed to be impossible to answer adequately, primarily due to learning being a long process. However, studying tendencies in their infancy can reveal where improvements can be made, especially when these are matched with results of previous research.

2 Theory

This chapter presents the main theoretical framework, called Activity Theory, that is used during analysis of the data obtained in the evaluation study. While the framework has no canonical practical application, it is valuable as a conceptual tool to guide reasoning and to understand activities as intricate systems, rather than as the sum of their parts. In doing so, the whole object-oriented process (past, present, and future) is considered and therefore the analysis is less at risk of deteriorating into a list of individual actions, an approach taken in many other frameworks in the field of human–computer interaction.

Section 2.2 discusses a set of heuristics for design of graphical user interfaces in computers. In the context of this thesis, the set of guidelines is used to analyse current design decisions and for making new ones with firm backing of theoretical results. The guidelines are used in practice and are taught at universities in courses in human–computer interaction. However, they have not gone without criticism, and we discuss their relevance in a modern software development and design project in Section 2.2.2, defending their position as useful when used in conjunction with other types of evaluation.

2.1 Activity Theory

Activity Theory (AT) is a theory of human activity, pioneered by Lev Vygotsky and his students in the early twentieth century in the Soviet Union. The unit of analysis is the *activity*, seen in a cultural and social context. This is an important distinction from other theories commonly used in human–computer interaction (HCI) research (such as the theory of Goals, Operators, Methods, and Selection Rules, GOMS [5]), where the unit of analysis is *actions* rather than whole activities. AT also deals with actions, but at a lower level and always in context of being part of an activity. Activities are considered to be directed toward an object and are mediated by tools of various kinds (languages, software, writing aids, etc). Development in any aspect related to the activity comes from contradictions within the activity system, for instance, the need for new rules governing the activity may emerge as the current situation is not satisfactory. The tools that mediate the activity are under constant development, and get their form and function from the requirements and uses of an older version of the tool.

The remainder of this section discusses the concepts introduced above in depth. In Section 2.1.1, the background and underlying concepts of AT are presented. Development via contradictions are discussed in Section 2.1.2. According to AT, activities can be viewed as hierarchical dynamic systems. This division is the topic of Section 2.1.3. Learning as personal development, rather than development of tools and of the activity itself, is discussed from an AT point of view in Section 2.1.4. Finally, Section 2.1.5 discusses some of the approaches used to apply AT in practice.

2.1.1 Principles of Activity Theory

AT originated in the beginning of the twentieth century in the Soviet Union, and was heavily influenced by the political and social climate of the time. It uses concepts found in dialectic material psychological schools of thought, that is, a Marxist view of societal and natural development. Development comes in cycles, and contradictions are seen as the force that drives change.

Vygotsky introduced AT as a psychological theory, aimed at understanding the mental capacities of a single human being [5]. However, AT holds that one cannot understand activity of an individual without the cultural and technical context in which it is carried out. Therefore, the unit of analysis is the *activity*, taken as a whole.

An activity, according to Vygotsky, has three fundamental characteristics: it is directed toward a material or ideal object, it is mediated by artifacts, and it is socially constituted within a cultural setting [5]. The artifact is also often referred to as a tool, and can be languages, technical equipment, and so forth. Vygotsky's system is often interpreted as a system consisting of Subject-Artifact-Object (this interpretation, however, is not quite true to the original, which dealt with stimulus and response, mediated by an artifact [10]).

Alexei Nikolaevich Leontiev, a student of Vygotsky, introduced the community into the equation, thus enabling AT to be used to consider complex situations such as Leontiev's "primeval collective hunting" example [22], which describes how one part of the hunter team makes a lot of noise to scare the game into the hands of the other part (which does the actual killing). Taken out of context, making noise is not related to the activity of hunting. In context, it is clear that it served a purpose that the community gained as a whole. However, Leontiev's model was based on Subject-Community-Object, disregarding the mediating artifact [5]. In other frameworks, such as Situated Action Models, this action would not be interpreted as part of a larger activity, although it is and should be understood as such [34].

Yrjö Engeström developed the ideas of Vygotsky and Leontiev further, and created an integrated model that is used by most AT researchers today. It is shown in Figure 2.1. Apart from the components already discussed (Vygotsky's Subject-Artifact-Object view and Leontiev's Subject-Community-Object view), it should be noted that the Engeström model has introduced others which were not present before. These are *rules* that govern the activity, the *division of labour* within the activity and the community and that the object produces some outcome, the object's *sense and meaning*. This marks the second generation of AT, according to Engeström [10].

The latest development of AT pushes the theory into its third generation, and it "...needs to develop conceptual tools to understand dialogue, multiple perspectives, and networks of interacting activity systems" [10, page 135]. Thus, the focus of study is not just one activity system, but at least two and the interaction between them. The outcome (object) of one system is the subject in the other, and there is also a transition phase between the two, where the object may be conceptually transformed into a new object. The object of the overall activity network system is a moving target, it cannot be understood by studying it as a conscious short-term goals as that would miss the "big picture".

As indicated by the first version of AT, mediating tools are a central concept. Tools are not seen as static artifacts — they evolve with the activity, and take on new forms and functions as the need arises. Knowledge is imbued into them, via a process referred to as *crystallization* [5]. In Section 2.1.3, the three levels of activities are discussed in depth. For an understanding of tool development, it is sufficient to intuitively understand that

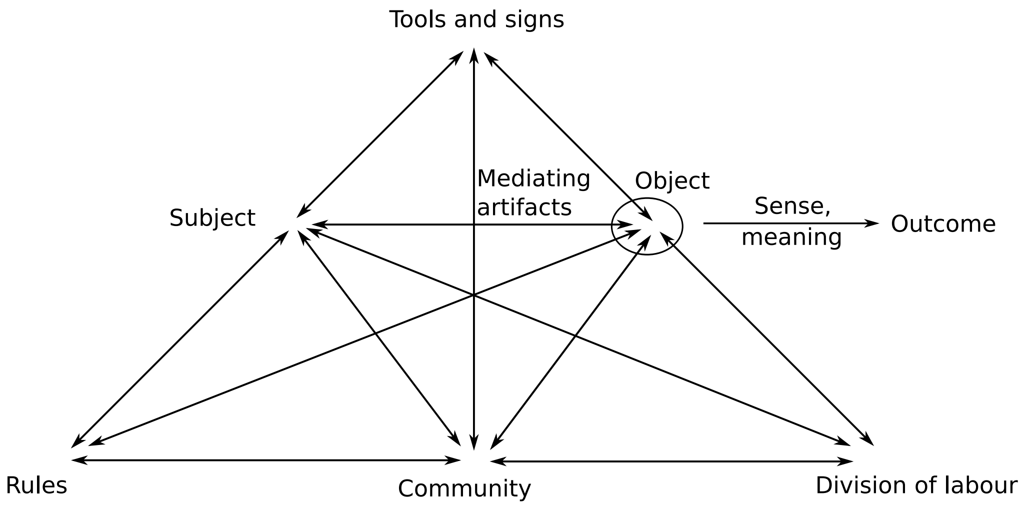


Figure 2.1: Engeström's model of activity systems, as shown in [10].

the lowest level of activities — acts that can be automated — are prime candidates for crystallization in the next generation of a tool. Aside from automatization, artifacts also crystallize knowledge by becoming representations of modes of acting in the given activity. A modern example are the spell checking facilities in computer word processors. Old versions of word processors required the user to hit a button to run the text through the spell checking program, but new versions do so on-the-fly throughout the entire writing process. Thus, an easily automated task for the user has been infused into the tool, and it serves the activity for which it was developed better.

Engeström argues that AT can be summarized with five principles [10]. The first principle is that of using the collective, artifact-mediated and object-oriented activity system, seen in a network of activity systems, as the *unit of analysis*. The second principle is that of *multi-voicedness in activity systems*. Participants in the activity — the community — are not copies of one another — they differ in opinions, experience, have different personal histories, and so forth. In addition, the activity system itself has multiple points of view expressed in the rules, artifacts and conventions as well as in the division of labour. All these factors need to be taken into account, as the activity system as a whole cannot be understood without them. The third principle is *historicity*, the historical context in which the activity system has been formed. Problems, division of labour, rules, the community and the other aspects of the activity system are functions of the history behind them, as activity systems are constantly developing in an evolutionary manner. The fourth principle is that *contradiction-driven development*, which is the topic of Section 2.1.2. Finally, the fifth principle concerns the *expansive transformations in activity systems*. Activity systems evolve in cycles, where the contradictions (internal or external) cause individuals to seek new and improved ways of carrying out the activity. An expansive transformation occurs when the object and motive of the activity undergo reconceptualization to incorporate a wider horizon of possibilities than the previous mode of the activity. Engeström describes this as a "collective journey through the *zone of proximal development* of the activity" [10] (emphasis in original). The zone of proximal development is a concept discussed further in Section 2.1.4.

2.1.2 Development via contradictions

AT considers the parts of an activity system, as shown in Figure 2.1, are in constant contradiction with each other. This view goes back to dialectical thinking (Hegel and Marx, among others), where dynamics are understood as solutions to internal antagonist contradictions within a system (discussed in [33]). Contradiction is not the same as conflict, and should not be regarded as a negative thing.

Contradictions are both internal and external in activity systems. When an activity changes via some external development, for instance when a new piece of technology is introduced, this can lead to contradictions with the division of labour, rules, and even what constitutes the community for the activity. Four classes of contradictions can be identified [5], and these are shown in Figure 2.2.

Engeström states that the primary contradiction in capitalism is between the use value and the exchange value of commodities [10]. This contradiction sets very basic and fundamental boundaries — a solution may not be developed until it is optimal, because that would be too costly with regard to the resources required for development. Simon's term *satisficing* [43] from the bounded rationality school of thought comes into play in these situations. It states, roughly, that one has to choose a solution that is satisfactory, rather than going for the optimal one, since the decision-making process itself consumes resources.

Secondary contradictions are between the corners of the activity system — easily understood as the contradictions between, for instance, the community and the division of labour in the activity. An example of this is the division of labour between physician (the community) when it comes to establishing the diagnosis for a patient. Can a general practitioner always ask a specialist? Are there rules that forbid this?

The tertiary contradictions are related to the development of the activity. The contradiction is between the current activity system and some other activity system that it could evolve into. The contradiction drives the development of the activity in a certain direction, producing a possible paradigm shift. In the case of DMSS-R, one such contradiction would be between the current system and one which is based on a tablet PC, which has implications not only for the graphical user interface, but for the activity of assisted dementia diagnosis as a whole.

Quaternary contradictions are between the activity and its neighbouring activities. As shown in Figure 2.2, there are several other systems one needs to take into consideration in addition to the currently studied activity system. The output of these systems become the input of this, and other, systems, therefore they are part of the greater context as well.

Activity systems are under constant development, due to contradictions and instability. New needs emerge, new pressure is put on the activity system, and the cultural and social context changes over time. The activity system affects the environment in which it is situated. Constantly searching for a product or solution that is satisficing with regard to some resource, activity systems never become static.

2.1.3 Levels of activities

As stated in the introduction of this section, AT is different from other theories in human–computer interaction such as GOMS [5] (Goals, Operations, Methods, Selection rules) in that it uses the activity as the unit of analysis, not merely the actions that constitutes it. This does not mean, however, that actions are disregarded in AT — rather, they are seen as part of activities and are studied as such.

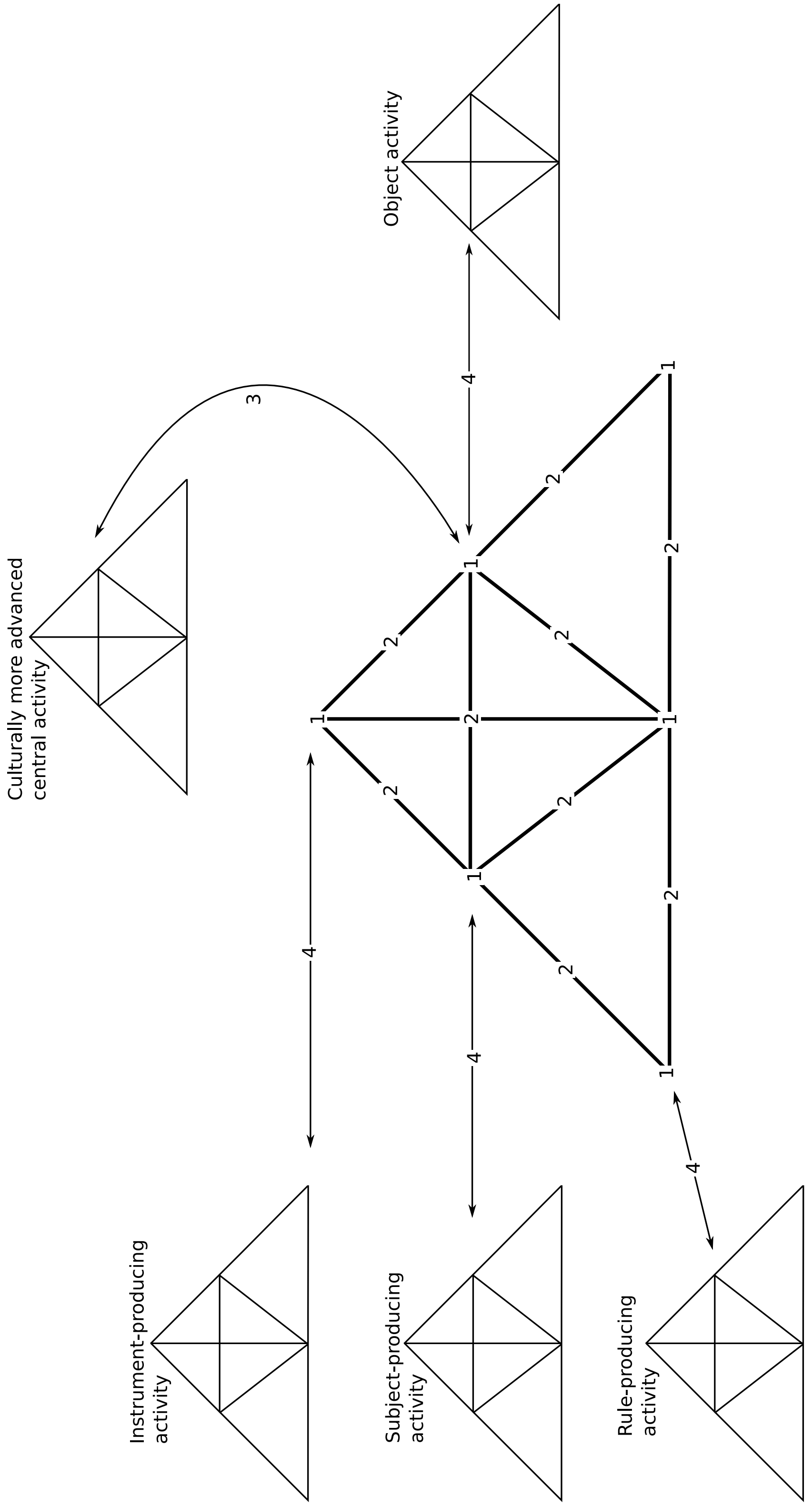

Object activity
Culturally more advanced central activity
Instrument-producing activity
Subject-producing activity
Rule-producing activity
1
2
3
4

Figure 2.2: Contradictions internal and external to the activity system as depicted in [5]. The numbers indicate the class of the contradiction.

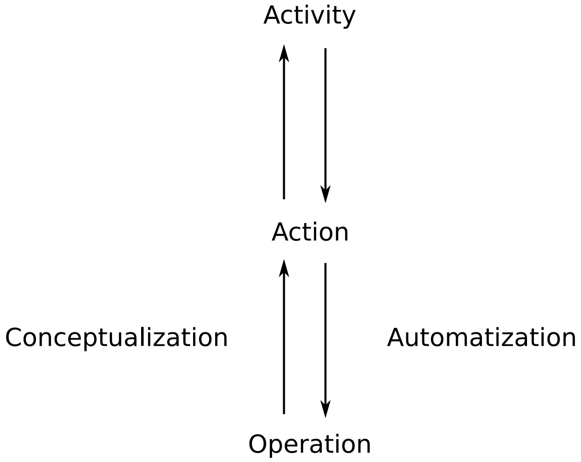


Figure 2.3: The dynamic system of levels of activity, as depicted in [5].

AT uses a three-tiered hierarchical approach to structuring activities. At the topmost level, *activities* are oriented toward a motive. Such motives are ideal or material, and the activity is carried out to fulfil some need. *Actions* are subordinated to activities, and are consciously formulated and carried out to — in some way that may not be immediately apparent, recall Section 2.1.1 — contribute to the activity. Actions consist of a number of *operations*, acts that can be carried out without giving them active thought.

Unlike some other theories (again, GOMS is an example of this [5], as is HTA [42], Hierarchical Task Analysis), the hierarchical division of activities is seen as a dynamic system in AT. People's behaviour in performing some activity differs according to their current context. In a stressful situation, acts that would otherwise be considered operations can require more conscious thought, and thus become actions. Likewise, actions can be elevated to the level of activities, given the right (or wrong, as it may be) circumstances. AT calls situations where an operation turns into an action *breakdowns.* A common source of breakdowns is not just stressful situations, but that the mediating tool somehow does not work in accordance with the subject's mental model of it. Such breakdowns are of great importance to the HCI field, since they are (usually) a sign of badly designed software [4]. Breakdowns are a special case of *focus shifts.* A focus shift simply occurs whenever a person shifts her focus to something else — intentionally or unintentionally (breakdown).

The opposite case, where an action becomes automated to such extent that it can be considered an operation, is called *conceptualisation.* Leontiev [22] uses an example to illustrate the dynamics in the activity hierarchy that is familiar to most people — learning how to drive. In the beginning, simple tasks such as shifting gears all require active thought — they are actions (the overall activity being to drive the car), all consisting of various operations such as hand and foot movements. As time progresses, these acts are conceptualised into operations — a driver can shift gears without having to actively think about how it is done. However, as stated above, if the gear shift for some reason stops working or the context changes, the operation may break down and become an action. Figure 2.3 shows the activity levels and the dynamic changes between them.

Kaptelinin argues that it is important to take the status of the behaviour in question into account, in order to understand and predict the behaviour of people in various situations [15]. One must ask whether the behaviour is oriented toward a motive, a goal, or actual conditions. Doing so makes classification of the activity easier, as asking this question lets us divide activities into the three levels shown in Figure 2.3. Whether the object of orientation for the activity is impelling in itself or is auxiliary is the difference between activities and actions. If the process is automated or not is the difference between operations and actions.

Another way of determining and classifying the levels of activity is discussed in [5] by Bertelsen and Bødker, based on work by Bærentsen and Trettvik. It uses the questions *why*, *what*, and *how* to separate the activity levels from each other. The activity levels are distinct from each other with regard to mental representation (what governs the activity level), what they realize and their level of description. This system is shown in Table 2.1. Note how context always remains an important factor, governing how operations may be carried out and affecting activities in the larger sense.

2.1.4 Learning and personal development

The dynamic nature of the levels of activity discussed in Section 2.1.3 has already shown that AT is a suitable framework for studying and understanding learning and personal development. Two important aspects of learning is that it takes time, and that it is a dynamic process. Just as in the case of the development of tools as described in Section 2.1.2, one must take the past into consideration to understand the present.

The process of conceptualization, turning an action into an operation, requires time and practice. While useful for describing how people learn how to perform manual tasks by sheer repetition, conceptualisation seems less applicable for intellectual tasks such as arithmetic. After all, while learning the multiplication table by heart certainly speeds up mathematical calculations, mere rote repetition does not teach the learner how to perform multiplication. In order to learn how to perform arithmetic, most were taught to count physical objects such as the fingers on one's hand. That is, a set of objects external to the mind of the learner. Once the learner can successfully count without resorting to counting the fingers, the concept has been *internalized*. At some later point, for instance when learning subtraction after having learnt and internalized addition, it may be required to revert to using external objects to represent the situation, the process of *externalization*. Computer use, decision support systems not excluded, is typically a case of externalization. The computer is instructed to handle large calculations or storing data that we simply cannot do ourselves. It is plain to see that learning, seen in this light, is a clear activity system, mediated by tools that allow us to perform externalization. The view that AT adopts, that the separation between mental (internal) and external representations is dynamic, departs from cognitivst approaches [5]. Cognition is integrated in the outward acts in which the individual engages.

The idea of internalization and externalization can be related to the theory of Distributed Cognition (DCog), which, in fact, has some similarities to AT [34]. DCog presents the view that knowledge is distributed among the individuals and artifacts, and is propagated between them. Thus, it is hoped that cognition can be studied at a system level rather than the level of a single individual. Hutchins, one of the largest proponents of DCog, illustrates this using a "cockpit system" (rather than "pilots operating in a cockpit"), where the success of the system is dependant not only on the pilots and their abilities, but also on the knowledge presented by various dials, lists and tools [14]. Reminiscent of AT, Hutchins argues that actions performed cannot be

Levels of activity	**Mental representation**	**Realizes**	**Level of description**	**Analytical question**
Activity	Motive (need) — not necessarily conscious, but may become conscious	Personality	The social and personal meaning of activity; its relation to motives and needs	Why?
Action	Goal — conscious	Activities (systems of actions organized to achieve goals)	Possible goals, critical goals, particularly relevant subgoals	What?
Operation	Condition of actions (structure of activity) — normally not conscious; only limited consciousness	Actions (chains of operations organized by goals and concrete conditions)	The concrete way of executing an action in accordance with the specific conditions surrounding the goal	How?

Table 2.1: Activity levels, and how to differentiate among them. As shown in [5].

understood without considering the entire system as a whole. One should, however, not be tempted to equate DCog and AT, as there are fundamental differences between them. DCog makes no real difference between a person and any other "node" in the system, their contribution equally valuable to the whole. AT recognizes the importance of mediating tools, but does not equate their contribution to that of the subject performing an activity. As mediating tools, smart systems such as DMSS-R, may be of particular interest in this case.

A concept of importance to pedagogy and developmental psychology is the *zone of proximal development* (ZPD). It can be understood as the zone between what a learner is already capable of and what can be learnt [5], usually via the help of a teacher of some kind. Vygotsky defined the ZPD as [50, p. 86]:

> the distance between the actual developmental level as determined by independent problem solving and the level of potential development as determined through problem solving under adult guidance, or in collaboration with more capable peers.

Extending this definition, which is mostly tailored to learning in children, one can add that "the collaboration with more capable peers" could also include using software to learn how to carry out some task.

It must be understood that the ZPD is of a dynamic nature, and that the amount of support needed by the learner to pass through it varies with experience. *Scaffolding* is a term used to describe the phenomenon of performance support and fading [49], however, it is not traditionally used in AT. An activity requires less support as experience increases, or, it requires support of a different kind. Too much or too little support has a negative hampering effect; the amount of support becoming a crutch or being insufficient, respectively. Learning is therefore usually best conducted in a problem-based fashion, where a big problem is presented and the learner obtains just the right amount of support from a teacher to handle the problem.

2.1.5 Practical application of AT

AT is a framework that is quite capable and well suited for describing human activity. However, there is no standard methodology for performing studies using it. This poses a problem to researchers, but it also gives the researchers some leeway in designing their studies — there is no set schedule or plan that has to be followed meticulously. Bødker argues that "[AT] helps structure analysis without totally prescribing what to look for" [4], and that AT constantly reminds the researcher of the importance of social and historical context of the activity that is being studied. Thus, one studies more than merely the actions and operations themselves (as one does in GOMS, for instance [5]).

There have been some attempts of guiding the thoughts of researchers when it comes to performing design and evaluation using AT. Kaptelinin et al. have developed a so-called "Activity checklist" for this purpose [16]. The checklist is divided in four parts: *means/ends* (hierarchical structure of activity), *environment* (object-orientedness), *learning/cognition/articulation* (externalization/internalization), and *development*. These categories are related to Kaptelinin's structuring of AT, where the principles are identified as being those presented as categories in the checklist, with the addition of a category for tool mediation [16].

To study the hierarchical structure of activity, Kaptelinin et al. state that in order to understand the use of any technology, one should start by identifying the goal of any action. After having done so, the action should be studied both upwards (see what activity it belongs to) and downwards (lower-level activities and the individual operations). The environment, or object-oriented aspect of activity systems, relates to the context of the activity as well as the objects that are in use during the activity. Given the importance of mediating tools in AT, understanding this aspect is crucial. The learning/cognition/articulation aspect is related to the object-oriented one, since some mediating tools enables the user to externalize certain cognitive features, while others enables the internalization of new knowledge. Finally, considering activity systems as dynamic and developing systems helps the researcher understand the history that shapes the activity in its current form and what the activity might be like in the future.

AT supplies us with tools for understanding and studying activities in humans, both from a distance (activities and systems of activities) and at a detailed level (actions and operations). Focusing on the details, as many do in interaction contexts (as, for instance, the aforementioned GOMS [5]), AT provides the concepts of the focus shift and the breakdown as targets for study. As stated in Section 2.1.3, focus shifts occur naturally and it is only the subset of them that are unintentional, the breakdowns, that are indicative of a possible interaction problem. Bødker provides the following checklist for studying focus shifts (the version presented here is abridged) [4]:

For each specific focus, ask:

- What is the purpose of the activity/actions for the user?
- Which object is focused on by the user? Where is this object located (in, through, or outside the computer application)?
- What is the instrument? Where is it located (in, through, or outside the computer application)?

For each focus shift, ask:

- From what focus/object to what?
- Is it a breakdown or a deliberate shift?
- What causes the shift: the physical, handling, or subject/object-directed aspects of the computer application?

The shift-causing aspects of the last point warrants further explanation. The physical aspect consists of how the user is using the tool physically, for instance, how one operates a mouse and a keyboard. If the user has not adapted or cannot adapt well to the physical requirements of the tool, the tool may become hard to use. For instance, a user with a disability which causes the user to enter data into a computer via some other means than the keyboard is very likely to experience data entry software differently than the majority of users. The handling aspects relate to the support for operations toward the computer application. If a breakdown occurs, the user must focus on the artifact, rather than carry out the operation it is supposed to mediate. Finally, the subject/object-directed aspect pertain the conditions for operations directed toward objects or subjects that are dealt with in or through the artifact itself.

Studying the focus shifts, and (using Bødker's checklist as presented above) determining which of them are deliberate and which are breakdowns, gives us valuable data in studying the interaction between the user and the tool. Bødker users a mapping technique [4] in two dimensions: one dimension contains the objects that the user focuses on, and the other the user's narrative of the situation, supplemented by annotations of the user's physical acting. By structuring the data in this way, focus shifts and thus the possible breakdowns become apparent and may be studied. An example of this can be seen in [5, page 319].

The zone of proximal development, ZPD (recall Section 2.1.4), may be studied and used to develop systems that aid the user in the learning process. Learning at an individual pace, where the challenge is just significant enough to be interesting, neither impossibly big nor boringly small, is best. This is used in some teaching software, such as Accelerated Reader [21], which presents texts of (supposedly) just the right level for the reader to comprehend and develop as a reader. Software that aims to be used for learning should thus be dynamic and adapt to the user just as the user adapts to the software, as static software will not be in the ZPD of many of its users for long.

2.2 Heuristic evaluation

Heuristic evaluation of computer interfaces is a method used by experts to evaluate systems from a usability point of view, without any required contact with users. The experts are assumed to have the users in mind during the entire evaluation, and be able to speak on their behalf. To do this successfully, the experts must have met the intended users and have a working understanding of their usage situation and routines. Heuristic evaluation is popular due to its quick and easy-to-understand nature, but has also been criticized for being very prone to human error and bias of the experts. Some counter-arguments are presented later in this section, as we argue that the usefulness of a quick analysis of an interface based on tried and tested heuristics is a cheap and efficient way of avoiding errors at an early stage of development.

While there are several sets of heuristics available, the ones that will be shown in this section are Nielsen's ten heuristics as they are presented in [36, 19]. The heuristics are shown in Section 2.2.1 and their use in modern interface design discussed in Section 2.2.2.

2.2.1 Nielsen's ten heuristics

Nielsen's ten heuristics, originally presented in the 1993 book *Usability Engineering*, remains one of the most cited and taught set of heuristics today. They are general in nature, and thus applicable in most situations. They have developed since they were first presented, and are here shown in their latest version, as in [36].

1. *Visibility of system status.* The system should always keep the user informed of any changes in status, and it should be done *within reasonable time.* Studies have shown that users will perceive a delay of even a mere second as disturbing [35] and that longer delays are thought-interrupting and cause them to want to focus on other things while the computer finishes its work. Keeping this in mind, all operations that can reasonably be thought of as finishing in more than a second need to inform the user of their status. A crucial aspect is also that the status change is *visible* — there is no point in updating an area of the screen in such a way that the user cannot see it without explicitly looking for a change to occur.

   The user will have some kind of goal state in mind, one that marks the completion of the tasks that needs to be carried out [44]. The system should provide the user with status updates so that it is plain to see that the goal state is being reached, step by step.

2. *Match between system and the real world.* To help users get familiar and comfortable with a system, it needs to use terms and concepts that belong to the domain of the user, rather than that of the system. For instance, a software for banks should use terms that are familiar in economics, and avoid programmer or

"system" terms. This goes for error messages as well — the user should not be greeted with cryptic error messages containing codes, but rather with a message that makes at least some sense to them.

This heuristic guideline is closely related to the study of mental models and metaphors. In understanding something new, we as humans often relate to other things that we already know — refining that knowledge when the metaphor breaks down and thus build our understanding of the new object. It is vital that any metaphors that are used in an interface is natural to the intended users, and that the metaphor is not broken down by the interface, causing confusion and irritation.

3. *User control and freedom.* Some users are explorers, whereas others are more comfortable with carefully planning their steps ahead of performing them [18]. Both groups, but the explorers in particular, have to be allowed to open various windows and forms in the system and there has to be clear exit signs that ensure the user that any action can be reverted to whatever state the system was in before. Changes that cannot be undone restrict freedom, and punish the user for attempting to learn the system on his or her own.

4. *Consistency and standards.* Systems should adhere to consistency and standards — consistency within themselves, and standards with regard to other systems on the same platform or of a similar type. A certain action should always have the same effect, so that it can be internalized with as little effort and as much benefit as possible. Widgets should be placed in a consistent manner, and their meaning should carry over across frames. Standards, such as shortcuts for how to perform the ubiquitous cut-copy-paste operations should be used in any new system as well, to let as much internalized knowledge be useful even in a new setting.

5. *Error prevention.* Humans make mistakes. For instance, we make spelling mistakes, misread labels instructing us to perform an action in a certain way, and generally just do things we are not intended to. A useful computer interface is built with this in mind, and actively tries to lessen the possibilities of making errors. Rather than text entry (which also requires tedious parsing on the back end), something like a drop-down menu or some other restricted but adequate form of entry should be used, if possible.

   Choosing the wrong item in such a situation is surely a type of error, but not one which can easily be solved via clever design. However, the risk can be greatly reduced if the words used to describe the choices are unambiguous and within the vocabulary of the user.

6. *Helping users recognise, diagnose, and recover from errors.* When errors occur, the system should inform the user of what has happened, help him or her avoid future errors by stating its reason, and help recover any data that might have been lost in the process. Software malfunction should ideally never cause the user to lose data.

7. *Recognition rather than recall.* All options, actions, and objects in the system should be made visible so that the user does not have to hunt for them. Related to the guideline about consistency and standards, the user will use software more easily if the interface is reminiscent of other interfaces they has seen and used than if they rely purely on recalling how to manipulate the software. Buttons

with familiar icons and text that state their purpose work toward the principle, cryptic commands entered at a prompt do not.

8. *Flexibility and efficiency of use.* Experienced users tend to prefer to use shortcuts in software, lest it feel restricting and awkward for them [44]. Ideally, these shortcuts can be customized to fit the individual user, as physical limitations may constrict their ability to use certain shortcuts (for instance, a one-handed user will have difficulty using a shortcut that spans across the entire keyboard). The system should not rely on shortcuts for all users, as novices tend to prefer to be guided through the usage experience [44].

9. *Aesthetic and minimalist design.* Information, in any form, be it text or graphic, should only be presented if it is relevant and/or often needed. A distracting and "busy" interface will cause the user to become reluctant to use the system. However, this must be balanced with the need for timely status updates.

10. *Help and documentation.* Help and documentation should always be available, and it should outline how to perform tasks in a step-by-step manner, as well as be searchable and easy to access and understand. There are mainly two uses for help and documentation: learning before attempting to use a certain feature in practice, and acting as a guide out of a situation that is beyond the current capabilities of the user. The first lays emphasis on the documentation being written in a style appropriate for new users who have little or no knowledge of the system. To support the second, the help system should ideally be context sensitive and provide help with whatever the user is trying to do — or, at the very least, let the user search for help topics using terms that are familiar to the user.

2.2.2 Discussion on heuristic evaluation

Heuristic evaluation has some strong points in favour of other types of evaluations: it is usually very fast (compared to other techniques, i.e. interviews or field studies), can keep costs down (due to requiring a minimum of test subjects), and the heuristics are based on years of research and experience in the field. The latter makes it possible to consider the current set of heuristics a tool with crystallized knowledge, in the Activity Theory sense (refer to Section 2.1.1). However, the method has some apparent drawbacks:

- the evaluation, performed by experts, is highly prone not only to bias (see Section 1.2.1) but also to over-analysis of details that are of little or no consequence for the intended users;
- results may be influenced by the knowledge of the expert; and
- heuristics may not always be applicable, and the general knowledge encoded or crystallized in them may in fact be wrong for the specific situation.

These criticisms are all valid, and pose a significant problem to studies that rely on heuristic evaluation. They will not be disputed here. However, we argue that as long as the study does not rely *solely* on heuristics, the method can be used as a complement to whatever other evaluation techniques are used.

In Scandinavian tradition, software development and interface design should be performed and created in a user-centered way [30]. As this is the goal for the DMSS project as a whole, applying something as inherently centered on the opinions and statements of experts as heuristic evaluation seems contradictory. We argue that heuristic evaluation is suitable to use, at least at (and possibly limited to) an early stage in the development of the user interface, to ensure that easily detected errors are found as early as possible.

All major computer desktop platforms have some sort of guidelines for how a well-adjusted application should behave to let the user experience the “look and feel” of the platform [7, 6, 12]. Ignoring these guidelines means that a user who is used to the platform will be confused when they are using the system’s interface. This should be avoided at all cost, and therefore it is prudent to ensure guideline/heuristic compatibility on the platform level as well.

3 Methods and materials

The methods used for data acquisition were developed in accordance with guidelines and models shown to be efficient and useful in qualitative studies in the medical domain. The framework for the evaluation study was laid out using the general DECIDE framework [42] and the more specific Guidelines for Best Evaluation Practises in Health Informatics (GEP-HI) [37].

In accordance with Action Research [2, 20], the study was conducted in a dynamic and reactive manner, considering each occasion a new iteration and adapting new iterations to the feedback provided during the previous iteration. DMSS-R itself underwent iterative development based on the results of the study, thus making it infeasible to conduct the study in a static way.

Section 3.1 describes the material used during the study. The methods, pluralised to indicate that there was no single method used during the entire study, are presented in Section 3.2, and the rationale behind them as well as a slight discussion is given in Section 3.3.

3.1 Material

This section describes the materials used for the evaluation study presented in the thesis. The material is summarised as follows:

- ten test subjects:
    - eight medical students, where two worked in pairs, and one could only complete half the study due to personal time constraints;
    - one intern (Swedish system: "ST"), who worked in psychology and had previous experience as a general practitioner. The intern did not actively work with dementia patients presently, but had done so to at least some extent during the time as a general practitioner — mostly by forwarding the patients to experts for a diagnosis; and
    - one domain expert, who is also involved with developing the system from the medical domain point of view (e.g. the medical terms and guidelines).
- three evaluation versions of the DMSS-R program; and
- recording equipment (digital video camera).

In addition, a questionnaire for determining attitudes toward the system and a set of questions asked during the actual test phase were used. These are presented in Section 3.2.

The largest group of test subjects used in the evaluation study consisted of volunteers among tenth semester medical students, studying at Umeå university. The tenth semester is devoted to geriatrics, neurology, social medicine, eyes, and the study of ears, nose and throat [29]. During the semester, groups of medical students spend two weeks time at the geriatrics ward (Geriatriskt centrum) in Umeå, one day of which at the department dealing with psychogeriatrics. During that day, they follow a doctor working

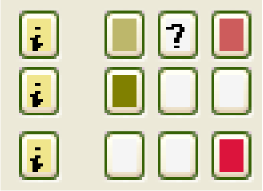


Figure 3.1: The green-red data input widget used in the first evaluated version of DMSS-R.

at the department and study and interact with dementia patients.[1] The other topics of the tenth semester are treated similarly, but are not of interest to this thesis.

Before the weeks of practical experience in the field, the medical students had lectures for two weeks at the start of the semester. Of these two weeks, two days were entirely devoted to selected topics in geriatrics. Between classes during these days, the students were asked to volunteer and sign up for the study.

The medical students were partitioned into groups for the semester, and during four two-week periods, the groups took turns visiting the geriatric ward. Thus, natural iterations for both refining DMSS-R and the study according to feedback were set by the schedule.

3.1.1 DMSS-R versions

Three different versions of DMSS-R were used, where the second and third were developed based on feedback from the study. The most obvious difference between the first two versions was the appearance of the data input widget,[2] whereas the third version featured a more drastically changed user interface. The first version of the data input widget is shown in Figure 3.1. Note the use of colours, green to the left and red to the right. The buttons were supposed to be mapped to indicate normality and presence of a pathological problem, respectively. This mapping was found to be problematic with questions regarding statements formulated in a way that required the "yes"-answer to be given by the red button and the "no"-answer by the green (i.e. questions asking for the presence of a certain disease). Therefore, the second version of the widget was revised as shown in Figure 3.2. Note that, in addition to opting for symbols rather than colour, there is also a header that explains the meaning of the buttons. The specifics of why the new look was chosen are given in Chapter 4. The third version (developed in May of 2008) used a data entry widget that was based on both previous versions and their respective strengths. Additionally, the user interface was changed to experiment with versions of the suggestions found in Chapter 5. For comparison, the widget used in the third version is shown in Figure 3.3.

The whole of DMSS-R itself will not be presented in this thesis, only the parts which warrant discussion. For more information on the system, see [24]. Also, it should be noted that any and all development regarding DMSS-R was done by people not directly involved with writing this thesis. However, changes were made based on suggestions

[1] Not all students have the same experience, as it is highly dependant on what the doctor needs to do for the day. There are no guarantees that all students do the same amount of interacting with patients, or any at all.

[2] Widget: a canonical term in human–computer interaction for "window gadget", an on-screen element in a graphical user interface.

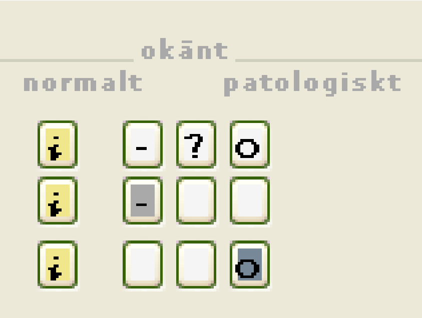


Figure 3.2: The symbol-based data input widget used in the second evaluated version of DMSS-R.

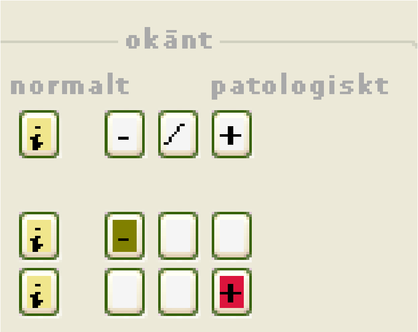


Figure 3.3: The third version of the data input widget, based on the two previous versions.

made in the process of analysing the data generated during the sessions of the evaluation study. These changes are discussed and motivated in the following chapters.

In addition to data input widget changes, the system was also updated in other ways between versions. Unfortunately, it is believed that changes in the underlying logic system also introduced programming bugs that affected the capabilities of the system to correctly diagnose patients, thus causing a reduction of data available in some aspects. These are discussed among the other results in Chaper 4.

3.2 Methods and Procedure

The eight medical students and the intern were used as test subjects in a three-part evaluation process. The parts were:

- a part were a fabricated patient's case was used for making a diagnosis using the system;
- a part where a an actual patient's case was used, using data that the test subject had remembered during his or her interaction with the patient during the day spent at the psychogeriatrics ward; and
- a written evaluation using a questionnaire comprised of seven questions.

The fabricated patient case was a typical case of the most common type of dementia: Alzheimer's disease. It was expected to use terms and concepts well within the knowledge of the test subjects. The text itself is available in Appendix A. The purpose of the fabricated patient case was primarily to supplement the limited experience and amount of patient cases of the test subjects. Another use of the patient case was that it helped investigate the medical reasoning skills of the test subjects. Because the correct diagnosis was known, as well as the data supporting it, incorrect reasoning could be spotted. In contrast, the actual patient cases were not known beforehand and were incomplete and the reasoning behind their diagnosis (if there was one) was not possible to verify.

During the first two sessions (involving three test subjects in total), the order of parts was as shown in the list above. During the remaining sessions, the order of the first two parts was changed. There were several reasons for this, the most important being that it was shown that focusing on a new patient case *and* a new system overloaded the cognitive resources and caused unnecessary errors. This change in method is supported by cognitive load theory (CLT) [49]. By using this feedback and inverting the order, the test subjects could focus on the new system with data that they had already internalised, resulting in more clearly shown breakdowns in the processes of investigation and use of the system. Due to time constraints, two sessions with medical students had to be kept short and only one of the parts could be conducted. For completeness, this lead to one session with only the actual patient case and another with only the fabricated patient case.

Both parts involving using the system were conducted as observation and interview sessions, complemented by “think-aloud”-techniques [44, 42], meaning that the test subject was instructed to verbally express any and all thoughts going through his or her head while using the system. The goal of the sessions was to find breakdowns (cf Section 2.1.3) in the interaction between human and computer, and to find other areas in need of improvement (i.e. related to supporting the reasoning or learning processes). A video camera was used to record 11 of 12 sessions for further analysis. The final session could not be recorded due to a sudden change of schedule.

The part with the fabricated patient's case was planned to be conducted as follows.

1. The test subject was shown a text, containing information about the patient. The text was written as a descriptive story, not structured in any particular way. The full text, in Swedish, is found in Appendix A. The test case includes elements from several actual patients, but is not based on any one patient, in order to anonymity. Worth noting is that the text is not complete — vital parts have been left out, making the text by itself inadequate for making a final diagnosis.

2. After being allowed to read the text for as long as the test subject chose, they were asked if there was any other information (i.e. test results) that they needed, and if so, what and why.

3. The test subject was asked if they already has a hypothetical diagnosis, and if so, what that was and why.

4. DMSS-R is used by the test subject to enter the data that has been collected from the textual description of the patient. The goal of this stage is to have the test subject enter data, and attempt to get a diagnosis from the system. Since the text did not contain all necessary information for making such a diagnosis, the system should at this point ask for the missing pieces of information.

5. The missing information is shown to the test subject as a way of simulating that further lab results have arrived, and once entered into DMSS-R, the processes of diagnosing the fabricated patient should ideally be over.

If the test subject fell silent for an extended period of time, perhaps confused by something in the system, they were reminded to keep talking. Given the questions that constitute the topic of the thesis, not only the usability aspects of the system were important, but also any and all thoughts about what additional tests to perform and if any new hypotheses were generated during the use of the system.

During the course of the observation session, unless the answers were provided implicitly during the think-aloud session or via observation, questions regarding the interaction with the system were asked. These questions were as follows (translated from Swedish):

1. (displaying main window) How do you start using the system? What are your first steps?

2. (displaying main window) Can you please explain your understanding of how the diagnosis tree works? When will it be updated, and in what way?

3. (displaying main window) Do you feel like you understand the purpose of every button on the display? Are the resulting actions clear to you?

4. (displaying main window) How may a diagnosis be obtained? What steps must be performed?

5. (displaying the "Status" data entry window) Is it clear what the current window relates to? Are the data entry fields easy to find, or would you like to rearrange them in any way? Do you understand how to use the data input widget to answer questions? Can you tell the functional and intended difference between regular

User	Version 1	Version 2	Version 3
Medical students	3	4	1
Psychology intern	0	0	1
Expert	1	0	2

Table 3.1: The number of sessions performed with the three different versions of DMSS-R.

Type of test case	Version 1	Version 2	Version 3
Fictive (medical students)	2	3	1
Fictive (psychology intern)	0	0	1
Actual (medical students)	2	4	0
Actual (psychology intern)	0	0	1
Actual (expert)	1	0	3

Table 3.2: The type of test cases performed with the three different versions of DMSS-R.

data entry widgets and the ones where severities can be entered, and if so, how are severities entered? How do you correct a mistake? If you feel like you need some help, can it be easily obtained?

The evaluation with the domain expert was carried out in a similar way to the sessions with the medical students and the intern. However, no fabricated test case was used. The expert was asked to use the system and think aloud with recent patient cases in mind. Three of these sessions were conducted, with different versions of DMSS-R. During the first session, the expert used the first version of the system for entering data from the first appointment with a patient. That same patient had a follow-up visit with the expert in May, and the new data of the patient was used during one of the later sessions with the third version of the system.

3.2.1 Number of test cases per version of DMSS-R

In Table 3.1, the number of conducted sessions per DMSS-R version is summarised. The table shows that the first and second version were mostly evaluated with medical students and that the third version was used more by the expert and the psychology intern. This is due to the late development date of the third version, dating to May 7th and the time constraints of the thesis work[3] and the low number of volunteers in the final round of geriatrics studies.

Table 3.2 shows the number and type of test cases per version of DMSS-R. The expert only used actual patient cases, whereas the other groups tested the system using a mix of roughly equal size of fabricated and actual patient cases.

In total, the evaluation study encompasses 7 fabricated and 11 actual patient cases, divided among 12 sessions. Note that the first actual patient case by the expert and the final one are the same patient, with new data added (resulting from a follow-up appointment with the expert).

[3] The deadline for the thesis was May 20th.

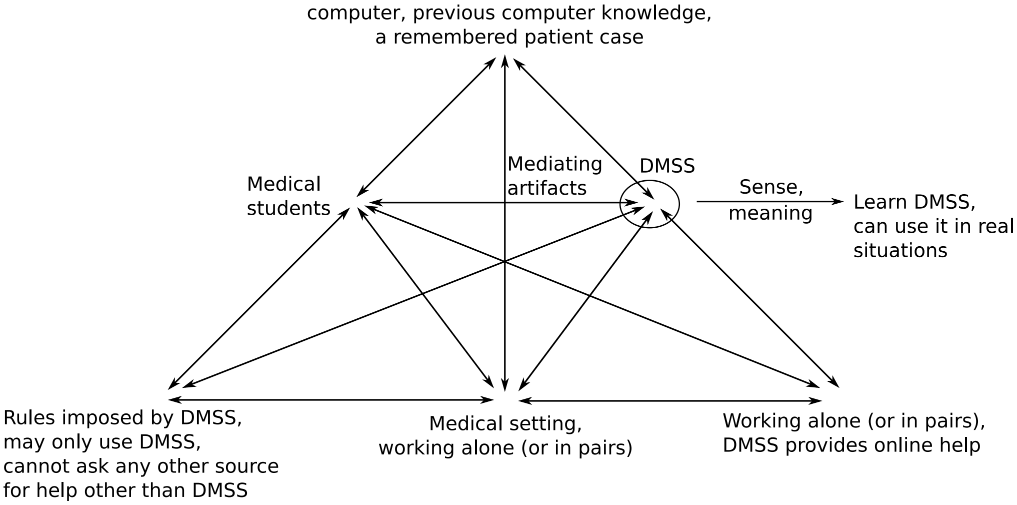

computer, previous computer knowledge, a remembered patient case
Medical students
Mediating artifacts
DMSS
Sense, meaning
Learn DMSS, can use it in real situations
Rules imposed by DMSS, may only use DMSS, cannot ask any other source for help other than DMSS
Medical setting, working alone (or in pairs)
Working alone (or in pairs), DMSS provides online help

Figure 3.4: Engeström model describing the first test phase for the evaluation of DMSS-R using medical students.

3.3 Method rationale and discussion

The data acquisition can be divided into three parts, with different goals in focus. For easy reference, we number the parts and consider the first one to be the one using the memorised patient case, the second to be the one using the fabricated patient, and the third to be the questionnaire. The main activity's object during the first part is letting the medical students get to know and understand the DMSS-R system. For this activity, the Engeström model in Figure 3.4 describes the activity as it was intended in its context.

In the second part, which consists of investigating the fabricated patient's case, the intention was for the focus to shift from learning DMSS-R to *using* DMSS-R in real situations. Thus, the corresponding Engeström model shown in Figure 3.5 deviates in key aspects when compared to Figure 3.4. As shown, the work setting is more realistic — the subject works with investigating an unknown disease as the goal, and may make full use of any internalised or otherwise available knowledge.

For the medical students, breakdowns are expected in both phases of the use of DMSS-R — the users are not experts, neither in their field nor as users of the system, and thus the setting contains many potential breakdown situations. By observation they can be studied and dealt with in a systematic manner. Supplementing the observations by using the "think-aloud" technique, it is possible to find not only breakdowns, but also regular (intended) focus shifts (provided the test subjects verbalise them).

The interview questions about the usability aspects of the system serve as gathering points for useful data of the graphical user interface. The questions were based on Nielsen's ten heuristics (refer to Section 2.2 and [36]). At the same time, they were intended to make the test subjects think about the different parts of the programs even further than they had during the use of the system. This was intended to aid in the learning process, since one must become aware of one's own tacit ideas and knowledge (which may be incorrect preconceptions) in order to explain them to someone else. Thus, once they have tried to explain various aspects of the system during the interview, the

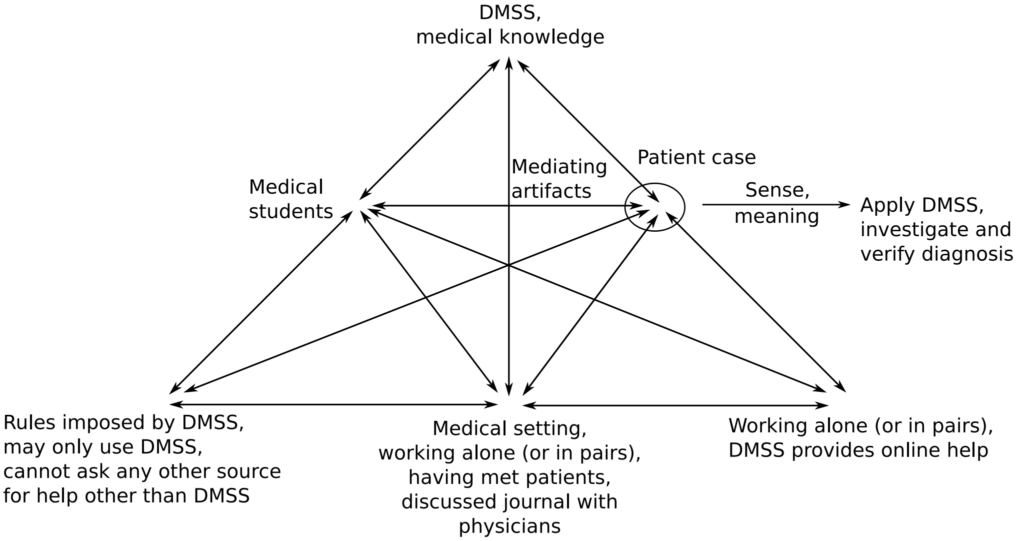


Figure 3.5: Engeström model describing the second test phase for the evaluation of DMSS-R using medical students.

usage of the program should be clearer.

In the second phase, the test subjects were asked to use the system with a fabricated patient case. The ambitious goal, as stated earlier, was not to teach them how to use the system, but rather to investigate the system as it is being used in a situation that requires medical reasoning and where the correct diagnosis is not known beforehand. Thus, it is shown whether the medical students, being non-experts in the field, learn something from the experience and to what extent they themselves think that they may be able to make use of a decision support system in their future line of work. During the test session, the intention was to let the test subject explore what it is like to learn the analysis domain using a tool.

The third phase containing the questionnaire gathers data in an easily comparable way. This data is used to investigate the attitudes of the medical students toward the DMSS-R system both as an evaluation of the system in its current state and for the future. In a recent study, it has been shown that medical decision support systems are not used as often as they could, although most physicians would agree that they are useful in principle [46]. The full list of reasons for this is beyond the scope of this thesis, but it seems reasonable to assume that at least part of the problem is due to lack of computer savvy — a problem that may diminish when young medical students and doctors who have grown up using computers enter the field. One of the major reasons why this type of system was not used according to the aforementioned study was that doctors felt that it disrupted the meeting with the patient (who, incidentally, didn't perceive the situation in that way) [46]. This factor has not been investigated in this study. By letting the test subject use the system behind closed doors, the problem has instead been circumvented. Therefore, the added stress factor of having to focus on both the patient and the system — which might be a contributing factor to why doctors feel that the use of decision support systems disrupt meetings, whereas the patients do not — is not present. One might be mislead into thinking, supported by the current view of usability testing (see criticism in [42]), that this yields data that cannot accurately portray the actual usage situation of the software. We claim that this is not the case — DMSS-R is not primarily intended for use during patient meetings, but rather for analysis of data afterwards. The software in the study by Toth–Pal was a reminder of sorts, one which acted as a cognitive aid for doctors in investigation situations [46], and, as such, was intended for use in another context than DMSS-R.

In addition to the medical students, evaluation sessions were conducted with a domain expert as well. The topic of the thesis is related to non-experts and the questions are to be answered from their point of view, but studying both their interaction with the system and that of an expert provides valuable insight by means of comparison. Testing with the domain expert allows for differentiating between that errors are made due to the medical students in the study being non-experts, and errors that are in common for both experts and non-experts. The assumption is that errors common for both groups are mostly of a user interface character, whereas the errors made exclusively by non-experts are assumed to be due to their lacking experience and expertise in the medical domain.

In contrast, the intern is thought to represent a middle ground between the medical students and the expert. The intern represents a more experienced physician, but with less domain knowledge in active memory. Both the expert and the medical students are very familiar with terms in psychogeriatrics because they actively worked with those topics during the evaluation study, the intern only to a lesser extent. Since the highly skilled expert is also very familiar with the system (due to being part of the development team), the intern who has never used the system before helps the differentiation between expert and non-expert even further.

As shown in Table 3.1, the third version was tested with a single last tenth semester medical student. This session helped gauge some of the implemented suggestions for future development presented in this thesis. The object of the session was to see if any of the features had the intended effect. The results of the session are presented in Chapter 7, as they are part of the basis for future work.

The think-aloud protocol comes with several limitations [23]. This is the reason it is only used to supplement the results from the observation and interview sessions. For instance, as users are engulfed in a problematic usage situation, they are less likely to think aloud and more likely to ponder the problem in silence. When they fall silent, they are encouraged to keep talking, but during the most crucial moments there may be complete silence. Also, not all focus shifts are verbalised: while scanning the screen for a certain field, for instance, the user might not state how this scanning is performed because it is several times faster to move the eyes than to actively state the intention of every single movement. This problem can be alleviated to some extent using eye-tracking camera devices [1], but it was not feasible for this study for financial and time budget reasons. The method was chosen, in spite of these limitations, since it is cheap and useful and at the same time generates a useful set of data.

In summary, the methods for finding troublesome aspects of DMSS-R are:

- observation;
- interviews;
- the think-aloud protocol;
- questionnaire on attitudes toward the system;
- interpretation of data using scientific theories and models; and
- theoretical results from various research areas (as presented in previous chapters), including relating to differences between non-experts and experts in a review-like fashion.

These methods are used together to triangulate problem areas of DMSS-R in need of improvement and to improve the validity of the results. That is, both results that differ or coincide are studied — the ones that differ because the source of the difference might require further study, and the ones that coincide because they indicate the presence of a particular problem. The validity of the results is ensured by corresponding results in the literature. Finally, the sessions with an expert and an intern help differentiate between the problems that were due to the medical students being non-experts, and the problems that were common for both non-experts and experts alike.

4 Results

The questions of Section 1.4 separate reasoning from learning by making them two areas of interest for the thesis. However, doing so for non-experts who use a system that helps them make decisions well beyond their current capabilities and authority is purely academic. In practise, the two processes cannot easily be separated — one never stops to learn, especially not while busy doing something that builds upon previous knowledge and experience such as making decisions or investigating a patient. This inseparable nature is one of the reasons why Activity Theory was chosen as the theoretical framework for analysing the data.

In order to structure the results in a presentable way, the following topics will be covered:

- learning how to perform the task of diagnosing a patient with support from the system; and
- learning how to use the system (including user interface usability aspects) and how the process of doing so may be influenced by the task at hand.

The usability concerns are the topic of Section 4.2, and the topic of learning how to use the system via the task at hand is discussed in Section 4.3. In the first section, we direct attention to those topics that relate to learning and performing the task of diagnosing a patient.

The evaluation study was, as stated in Chapter 3, split in two parts. The reason was that attempting to both concentrate on the fabricated patient case and learning the system appeared too demanding for the test subjects. Changing the study so that the test subjects first worked with a memorised patient case was highly successful. In doing so, the cognitive load lessened tremendously and the test subjects could focus more attention on learning the system. Due to the higher level of concentration of not having to check the textual description of the patient case for answering every data entry point, the focus shifts that did occur were easier to identify and deal with in a systematic manner during analysis.

4.1 Reasoning and medical investigation

Diagnosing a patient, first and foremost, requires vast amounts of medical knowledge. Based on this knowledge and results of tests that relate the knowledge to the particular patient, hypotheses can be made. These are refined as more data (test results) is revealed, and the set of generated hypotheses is reduced to the most likely ones, forming the final diagnosis, once sufficient evidence has been discovered to assure the physician that the diagnosis is correct. We shall study these steps, and how the system relates to them.

In this section, we focus on the results of the study with the fabricated patient case. The intention of the phase of the study where an actual patient was used was to let the test subjects get acquainted with the system. The phase is not used to study medical reasoning, because the underlying data was so diverse (and in many cases incomplete) that it would not be possible to do so without the help of an expert (and access to the data itself). However, using the fabricated patient case, both the amount of data at the disposal of the test subjects and the correct diagnosis is known.

In presenting the results of the evaluation study using the fabricated test patient, we recall from Section 3.2 that the overall intended activity system is as shown in Figure 3.5. The medical activities, in Activity Theory terms, are to establish a correct diagnosis for the fabricated patient and to investigate the correctness of the diagnosis of some other patient. Considering the fabricated patient's case, the actions related to establishing the diagnosis are:

- to study the text describing the patient's case;
- to interpret the description into data and evidence;
- to formulate a set of hypotheses, based on the data and evidence and the relationship between them;
- to consider an additional set of tests that would gather pertinent data, in support of or contradiction to these hypotheses;
- to verify if any of these hypotheses can be shown to be correct; and
- to make a final diagnosis.

The operations in this process are first and foremost internal cognitive processes, but also encompass the operations needed for externalisation (such as writing down ideas or highlighting important parts in the description of the patient) and internalisation (reading the text repeatedly with the highlights in place). Of particular interest are the steps that can be supported by DMSS-R, i.e. generation, refinement, and verification of hypotheses.

4.1.1 Hypothesis generation

The test subjects did precisely what has been documented on non-experts and hypothesis generation. They, as a group, generated several sets of various sizes of hypotheses on what type of cognitive disorder the fabricated patient was suffering from. All included the correct hypothesis (Alzheimer's disease) in the set, but to an expert, the case is perfectly clearly a typical case of Alzheimer's disease. This agrees with the result of Patel et al. [40].

Furthermore, the (too) large set of hypotheses also made the test subjects request a large number of additional tests. An expert, knowing the guidelines for determining a state of Alzheimer's disease, would require no extra tests than those that were given. This, too, agrees with findings of Patel et al. [40] Some test subjects requested clarification on some points (the time perspective being a common one), which, although documented, were arguably not as clearly stated as possible. These were not counted toward the total number of additional tests that were requested.

Table 4.1 show the hypotheses generated by the medical students and the intern who worked with the fabricated patient case as well as the requested extra tests for investigation of the patient. For anonymity, the numbers for marking the sessions are devoid of all meaning and used only for easy reference in conjunction with Table 4.2 showing the diagnoses obtained by the test subjects.

Session	**Hypotheses**	**Requested additional tests**
1	Alzheimer's disease	None
2	Some type of dementia	None
3	Maybe dementia? Not vascular dementia, but maybe Alzheimer's disease?	Neurological tests, investigate other causes like focal neurological problems, paralysis. CT-scan. Current medication? Ventricular fluid tests.
4	Alzheimer's disease? Maybe frontotemporal dementia?	Neuroradiology
5	Could be many things — stress and aggression are indicators of frontotemporal dementia (although it is a fairly uncommon type of dementia), but it could also be frustration due to suffering from Alzheimer's disease. Preference for frontotemporal dementia.	CT, lab tests (homocystein, folate, ...), ventricular fluid tests
6	Alzheimer's disease, but cannot rule any other type out	Blood tests and tests for Parkinson's disease.
7	Could be dementia, but not necessarily. Could be Alzheimer's disease or a microvascular dementia type. Vascular dementia likely.	None

Table 4.1: Sets of hypotheses generated and additional tests requested by the medical students and the intern.

The evaluation study has confirmed that not only do the non-experts generate a larger set of hypotheses than necessary, they also request tests that are not pertinent to determining the true cause of the cognitive disability. This result is hardly surprising as it has been documented in the literature. The task for a decision support system such as DMSS-R is therefore to limit the potential damage and cost that arise in these situations. In particular, it is important for such a system to help the user refine hypotheses along the reasoning process, pruning away those that are not relevant.

4.1.2 Hypotheses refinement

Once the initial set of hypotheses had been generated, the test subjects were instructed to start using the system. They entered the data as correctly as possible according to their understanding of it, and (due to omissions) the system requested that they enter more data after the users had used the analysis function. The test subjects were asked if they had refined their initial set of hypotheses due to the questions asked by the system, as it was believed that the questions would spur the users into thinking about relevant topics that they had not thought of before. For instance, since a steady decline in cognitive ability is a trait common for Alzheimer's disease, having the system ask about the time aspects of the disease might have guided the test subjects toward remembering that particular fact and its importance. However, this process was not expressed by nor observed in the test subjects.

One test subject could, after some time, rule out a vascular dementia — but had already almost done so in the beginning, and did not attribute the change in judgement to having used the system. Rather, it was due to having thought about the case for some additional time.

We have found that none of the test subjects working with the fabricated patient case, to any measurable extent, refined their set of hypotheses as an effect of using the system. Neither did they refine their set of additional requested tests. Due to being non-experts, and the limited time for the sessions, it is possible that the test subjects did not actually reach the step where they start refining their set of hypotheses. This shows that the system could offer more support in this area, as it did not appear to reach the ZPD of its users and help them advance further in the process.

4.1.3 Verifying the diagnosis

In line with the lack of hypothesis refinement exhibited by the test subjects, it comes as no surprise that the test subjects were also quite trusting of the system. A programming error, discovered after the sessions, caused Alzheimer's disease to be misdiagnosed (in fact, no diagnosis could be given) by the system in three of seven sessions. In the cases where the system actually reached a diagnosis, the test subjects were convinced that the diagnosis was correct. Since the system can only work with the data provided to it from the user, the diagnosis is only as correct as the data it is based on. In attempting to establish a diagnosis for the fabricated test case, where the correct diagnosis was Alzheimer's disease, only one of the test subjects received the sought result. The system suggested "Mild cognitive impairment" during two of the sessions, which indicates the presence of cognitive impairment but not a state of dementia. Another was informed that the diagnosis was frontotemporal dementia (FTD), whereas the remaining three were not provided with a diagnosis at all, indicating the presence of some software bug. Table 4.2 shows the diagnoses that the system provided the test subjects with. The

Session	Diagnosis	Version of DMSS-R
1	Alzheimer's disease	1
2	Mild cognitive impairment	3
3	None given, bug?	1
4	None given, bug?	2
5	Frontotemporal dementia	2
6	None given, bug?	2
7	Mild cognitive impairment	3

Table 4.2: The diagnoses given to test subjects by the system and the version of the system.

numbers used in the left margin correspond to those in Table 4.1.

The bug was thought to have been introduced by version 2, but as shown, test subject 3 used version 1 and was also subject to some programming error. DMSS-R should, barring programming errors, always either reach a suggestion for a diagnosis *or* keep asking for data until it can do so. Table 4.2 also shows that both users of the third version of DMSS-R were suggested "Mild cognitive impairment" (MCI) as the diagnosis. MCI is not a state of dementia as the impairment is too mild; unless this is a sign of another program bug, this is perhaps an indication of difficulties concerning entering the correct severities of the problems of the patient (discussed at length in the upcoming section). Also, we note the effects of possible bias in the case of test subject 5. The set of hypotheses included frontotemporal dementia as a likely candidate, and the data entered into the system suggested that this was indeed the case. Since the correct diagnosis was an obvious case of Alzheimer's disease, this may be seen as an example of confirmation bias. It is also possible that the evidence in the patient case was simply interpreted wrong, or entered erroneously, because Alzheimer's disease and FTD are similar and in certain cases may be differentiated by a single feature (visio-spatial ability). This similarity requires that there is no X-ray evidence indicating the contrary, and that the symptoms are of the set common to both diseases.

The effect of trusting the system may be explained by the fact that the test subjects were non-experts. Daley found that novice (her term for referring to non-experts) learning processes are affected by fear, mistakes, and a need for validation [8]. Due to this, they will more readily accept what others tell them and soak up information. DMSS-R is seen as having more authority, which means that the non-experts imbue a level of trust in the system similar to that invested in teacher.

Another probable explanation, although the data is not large enough to give conclusive evidence, is that the test subjects that possessed the most computer knowledge also exhibited the highest level of trust toward the system. During informal interviews, it was established which of the test subjects used computers to a large extent and which did not. The results seem to indicate that if a person is comfortable with using computers, the person also trusts the computer to function correctly. People who are more cautious with computers displayed more mistrust and did not as quickly accept the diagnosis the system had made.

Since the analysis conducted by the system is only as good as the data it has been provided with, and since the non-expert users may tend to use the system as a trusted source of information, it is crucial that any errors due to faulty data are clearly displayed. If not, the non-expert user has no chance of finding why the system would suggest a certain diagnosis that might go against what the internalised medical guidelines say.

The remaining cases, using actual patient data, will be analysed in an ongoing study. Due to the higher level of medical knowledge required to do so, they have not been covered here. They do, however, provide interesting data — in particular the case of the returning patient, where the system could play an active part in discerning the type of dementia.

4.2 Graphical user interface

To find possible areas of improvement in the graphical user interface (GUI) of DMSS-R, the technique outlined in Section 2.1.5 was used. To recapitulate, it states that by studying all focus shifts that occur during a session of usage, the involuntary focus shifts can be identified. These are referred to as breakdowns, and the causes for such situations are problem areas in need of improvement. The topic of usability is very important, as it has been shown clearly that one of the major sources of error in medicine today is faults in the interface between humans and technology [17].

4.2.1 Data input widget breakdowns

One of the main tasks carried out while using DMSS-R, if not *the* most important from the point of view of the user, is data entry. Once data has been entered, the system can perform its task of acting as a decision support system. Most of the interaction with the system is via the data input widget. Previous versions of the widget were based on drop-down menus [24], but these proved to be problematic for various reasons.

The GUI of DMSS-R changed twice during the evaluation study, as stated in Section 3.1. The first two sessions with medical students (using the first version of the system) showed that the data input widget posed a problem, and was a near-constant source of breakdowns during one of these sessions. The primary reason for this was that the widget was considered to be inconsistent, and because it used culturally ingrained symbols of colour in ways that were hard to understand. The problem was that the widget had a green button, and a red button — colours that have meanings such as "ok"/"yes" and "not ok"/"no", respectively, in our culture. The idea behind the choice of colours was to exploit the fact that green could be used for indicating that something was normal, and that red was abnormal. This, too, was considered a problem in one of the sessions, but the major breakdown was caused by questions posed in a "yes or no" way. Due to cultural influence, one would assume that the affirmative answer would be correctly entered via the green button. However, since the question was asking for the presence of a problem, the button that correctly indicated the abnormal/problematic case was the red one. This was not indicated by any clues in the GUI, and had to be stated by the evaluator. A test subject stated that he/she did not know what the data he/she had just entered into the system actually meant: if it was the affirmative or negative answer.

In light of the problems with the coloured data input widget, a new one was created for the second version of DMSS-R. It relied on symbols and shades of grey rather than colours. In addition to these changes, headings that stated what the buttons meant were added above the columns of buttons. The reasons behind these choices were:

- red and green might be a problem for colour-blind users, and the shades of grey would be more readily apparent to them;
- symbols might convey the message more easily than mere colours; and

– an explanatory heading over the column of buttons would help in answering the questions posed by the system, even if the choice of button might be experienced as counter-intuitive.

The symbols used in the second version of the system on the buttons were "-" and "o". The former being the one indicating negative test results (what patients see as a positive thing, that is, the normal state) and the latter indicating the presence of a problem. The choice of "o" rather than "+" might seem surprising given the intended positive meaning, but one should keep in mind that DMSS-R is developed for the Asian market as well [24], where "o" is in fact the positive symbol. The intention was to see if "o" could be interpreted as a positive symbol, regardless of cultural influence. Additionally, the "o" symbol is more readily told apart from the "-" symbol than the "+" when presented on a computer screen.

The results of this change were mixed. Via observation and through comments made from a test subject, it became apparent that the symbols did not provide a clear clue for the users. When asked about the symbols, and if they were useful, answers indicated that the symbols were hardly noticed — rather, the headings of the button columns were used to find the appropriate button. Once the headings were internalised and the use of the buttons were conceptualised, it became more a matter of thinking in terms of "right" and "left" (the "-" button was on the left, and the "o" on the right) than finding the appropriate symbol, according to test subjects. This gives rise to the following somewhat contradictory result:

– removing the culturally significant colours of red and green and replacing them by symbols made the user interface less intuitive; and
– due to being less intuitive, the risk of making input errors due to interpreting the colours incorrectly was reduced (increasing the reliance on the headings rather than the quick mapping the colours provided).

In fact, one test subject showed such frustration with the new symbol-oriented button labelling system that he/she expressed a wish for coloured buttons — suggesting red and green as suitable colours.

The second result is a case of *deceptive affordance*. Affordance is a term used in the literature for the property of things whose correct usage is immediately apparent upon seeing them [42]. Deceptive affordance, then, is the case where something exudes something similar to affordance — but signals an incorrect usage. Since the user, due to relying on the signals, will happily keep making mistakes, deceptive affordance is a big problem from a human–computer interaction point of view.

The third version of the system reverted back to using red and green, in addition to symbols and headings — however, exchanging "o" for "+" and "?" for "/". Unlike the first version, the colours were not shown until the user had clicked on either of the buttons for positive or negative. The reasons for this decision was that the colours should not be a constant distraction, but mark which questions had been answered and in what way. The choice of symbols was due to feedback from test subjects — the "o" symbol did not appear to have any meaning, and "?" was mistaken for "help" (as is common in user interfaces). Results from the sessions conducted with the third version suggest that the plus and minus symbols are not as intuitive as desired, but that once their meaning has been pointed out, they serve their purpose. Further studies are needed to establish whether the symbols should be used or replaced, and if so, by what.

To use Nielsen's terms (cf. Section 2.2), the first version of the widget was subject to problems of inconsistency and broken standards. In the second, these were exchanged for a new set of problems stemming from poor match between the system and the real world, due to using symbols without inherit meaning. The third version of the widget was designed, based on the strengths of both previous designs. Further evaluation is required to see if these changes are as sound as the limited amount of testing with them seem to indicate.

Common for both the coloured and the symbol-oriented version of the data input widget were the following:

- The test subjects all saw the set of widgets as a list, where the appropriate answer should be ticked. This includes the "no data" button (marked "?" or "/", depending on DMSS-R version), indicating that no data is available. The fact that the buttons resemble a list is, in itself, not a problem. It is related to the mental model that was intended, but clicking the buttons is a time consuming process.
- Clicking the "no data" button did not give a visual feedback of successful operation, if it was clicked while the widget as a whole was in the original state. This caused the test subjects to click it repeatedly until they concluded that the lack of feedback was intentional and the system was working correctly.

The problem of users feeling the need to click the "no data" button even though it has no effect can both be considered either as a problem or as a usage pattern. It is a problem in that it is time consuming, but it may also be a simple trick the users use to mark the end of their consideration of a question. Once the question has been read, the user must ponder the correct answer and, whatever that answer is, click a button. This, however, requires that the user actually reads the text first and then clicks a button before moving on. Otherwise, clicking the "no data" buttons is simply a waste of time — an action performed only due to thinking that the system requires the user to do so. The latter situation came up in several sessions, in particular, a test subject stated that he/she did not have any data for the entire right side of a screen, only to proceeded clicking the "no data" buttons until all rows had been clicked.

The second problem related to the button in question is that it gives no feedback once it has been clicked, resulting in repeated clicks. Considering the users who mistakenly view clicking the button as a required action, making the button emit some feedback when it is clicked might deepen the misconception that it must be clicked. Nielsen's first rule of visibility of system status states that changes should be reported to the user in a timely manner. This can also be interpreted as a rule dictating that *only* changes should be reported, thus making the lack of feedback the appropriate choice. Any confusion that arises will have to pass.

Entering levels of severity

The data input widget serves not only as the widget for entering binary or tertiary data (considering "no data" as a value), but for any range. A special version of the widget is used for this purpose, and it is shown in Figure 4.1. To use the widget for entering severities, one should click the button indicating the presence of a problem several times until the desired value is shown in the box to the right. No other forms of data entry may be used (i.e., placing the cursor in the neighbouring box and typing the appropriate number). The latter proved to be a problem to test subjects who tried to do so immediately upon hearing that the box was used for entering severities.

The severity input widget, too, underwent some changes in addition to the colour/symbol change as described in the previous section. The first version (topmost in Figure 4.1) would default to the value "?" for the first click on the red button. The difference between clicking the "no data" button and just clicking the red button once was unclear

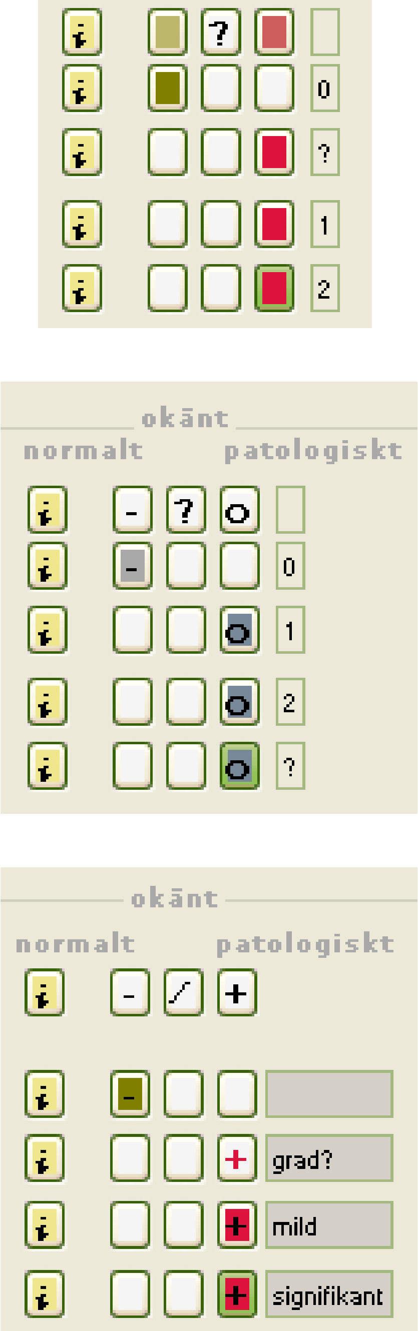


Figure 4.1: Widgets used for entering levels of severity in DMSS-R. The three different versions of the widget are shown from top to bottom starting with the first version.

(in fact, there was none). This was changed in the second version, so a click on the "o" button would set the severity value to 1. The rationale was that if the user clicks the button, it should reasonably mean that the symptom is at least mild. In the third version, this was changed (as evident in Figure 4.1) so that the first click indicates the presence of a problem, but to an unknown degree. This new uncertain value was placed "first" (requiring the least amount of clicks on the button) because it makes sense that subsequent clicks should increase the level of severity.

However, the widget, regardless of using colours or symbols, has proved to be a usability problem and a cause of breakdowns for the test subjects. No test subject paid any attention to the difference between the severity input widget and the regular data input widget. One of the test subjects stated that they thought that the system might fill in the values itself, based on whatever else had been filled in. But no one explored the widget, and all had to be told what it was used for. The functionality of the widget is explained in the online help system, but not a single test subject used the help system voluntarily to find out about the system. One was instructed to do so, and used the widget correctly afterwards.

In the first two versions, severities could be entered on a scale from 0–2 (normal, mild or severe) and in the third, these numbers were replaced with text labels. Text labels were chosen because test subjects did not understand what the numbers really meant. Just as with the case of answering questions with either a "yes" or a "no", test subjects found it hard to know precisely where to draw the line between mild and severe deviations from the norm. Again, this was not a problem to the expert who had a much clearer concept of how to discern a mild from a severe impairment of some function. Thus, for non-experts to use the system in an adequate way, there is a need for supporting their more vague understanding of the terms and concepts used in the system. The third version of the system replaced the numbers with text in an attempt to call more attention to what had been entered. Observations of the change show that test subjects identified the entered level of severity quicker (once entered), but that the change did not help them understand how to use the button correctly.

4.2.2 Feedback-related breakdowns

DMSS-R uses several techniques for informing the user that something on-screen has changed, the most important of which is a portion of the main screen with text that changes depending on the information that should be presented to the user. However, the evaluation study has shown that the test subjects experienced breakdowns related to the amount of feedback given and that they did not notice that the text changed (and thus missed the main channel for obtaining feedback). In no particular order, these feedback-related problems were:

– The text that indicates the current stage of the analysis process of the system, shown in the main window of the application, is not paid attention to. The text is black on a grey background, and presented without headings or anything else that might quickly indicate that it has changed. Test subjects assumed it remained static throughout the session, and did not notice what it said until it was pointed out to them that it had changed.
– Since the text in the main window is not given enough attention, the test subjects felt that there was no indication of what they were supposed to do or what the system was busy doing.
– DMSS-R can be used in one of two ways. One way is to simply hit the "Analysis"

button immediately, and fill in whatever features the system requires, the other is to fill in as much data as possible first and then request an analysis. The majority of test subjects (four of six) preferred the latter approach. However, when they had filled in all data they had gathered, DMSS-R required them to click the button several times to make the system advance. This caused confusion and frustration, especially among the test subjects who at that point read the aforementioned descriptive text in the main window, instructing them to enter data (although no additional data was needed at that time). This was changed for the second version, so that the system attempts to take the analysis as far as possible, based on the data at its disposal.

- Upon using the “Analysis” button, the user is given no feedback at all that the system is working. On a fast computer, the processing should not take long, but if the system cannot continue for some error-related reason, the user will not be informed that anything has happened. This causes feelings of being annoyed with the system. This is a problem with the error handling in the system, because the system should never enter such a stopping state.
- Once analysis has been able to suggest a diagnosis, the user (if the system could determine a type of dementia) will be presented with a screen for suggesting suitable intervention for the patient as a part of the analysis process. This is a problem with consistency, since the button marked “Analysis” starts to perform a new action once the system has been able to suggest a diagnosis.

When DMSS-R during analysis determines, using the guidelines encoded in the logic back-end system, that some data is required and missing, the text next to the data input widget for that data is marked with red colour. Test subjects found this indicator hard to find, often scanning the entire screen several times before locating the red text. One can only assume that a colour-blind person would have added difficulties in doing so.[1] Since DMSS-R only performs its analysis of the data when the “Analysis” button has been pressed, it is quite possible to initiate a round of analysis, be informed of missing data and rectify the situation, and then initiate another round of analysis. In some cases, the red markings were still present in subsequent rounds of analysis, even though the data had been entered previously. This caused the test subjects to ask if the system “thought the data was wrong” or if there was some other problem.

Inconsistency in these red markings were also a cause of breakdowns. The inconsistency was evident in the “Status” and “Heteroanamnesis” windows, where many important data entry points (“core features” in DMSS-R [24]) reside. There, the markings worked differently in the first two versions. Instead of marking the text next to the data input widgets, the frame surrounding the widgets (including the header for the frame) is highlighted. The test subjects found this confusing and hard to notice — they were searching for marked rows, not paying attention to the frame and its header. This was changed in the third version, so that the individual rows with missing data entries were marked.

DMSS-R should provide timely feedback that the analysis is working, has finished and, if necessary, clearly state that no result could be obtained. Nielsen’s guidelines and previous research show that users expect systems to respond within a highly limited time frame, and therefore DMSS-R should provide such feedback. Additionally, the “Analysis” button should be disabled in situations when it cannot reasonably be used for anything.

[1] Remark: test subjects were not asked if they were colour-blind.

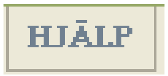


Figure 4.2: Help button, as shown in the data entry frames.

The breakdowns presented in this subsection are all symptoms of not adhering to Nielsen's heuristics. While the system does provide feedback, the feedback is not factually correct in all cases (asking for additional data even though none is needed). Additionally, the heuristics concerning consistency and standards, helping users recognise errors, and a combination of the lacking match between the system and the real world and matters of user control and freedom were broken to some extent. The latter combination is due to the fact that the system seems better suited for one of the usage patterns than the other (filling in only the marked values versus filling in as much as possible and then requesting an analysis).

4.2.3 Breakdowns due to unclarity

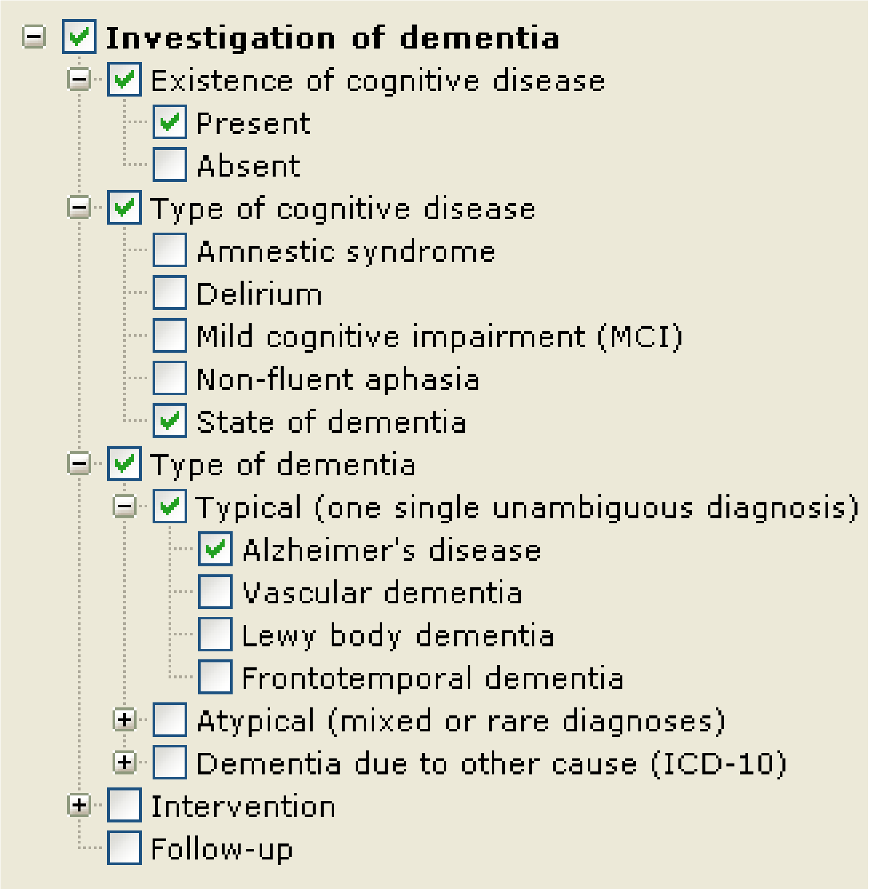
It stands to reason that most, if not all, breakdowns in the graphical user interface of DMSS-R are due to unclarities in some sense. Thus, the previously discussed breakdowns could have been presented in this section and its header. However, this section is devoted to the various other aspects of DMSS-R that have been identified as particularly troublesome for the test subjects.

Virtually none of the test persons attempted to get online help from the system. Those who viewed it only did so once instructed that it might contain useful clues to aid in a particular situation. This is indeed a problem, as it shows that the test subjects (most young and most fairly used to working with computers) incorrectly assumed that they would be able to figure out how the software worked without reading the documentation. The experience from the industry indicates that many do not read documentation at all, preferring to be guided either with a demo of some sort or a very short and poignant list of steps. The aspects of the system that have been found to be troublesome, including the data input widget, are explained briefly in the online help and perusing it would likely have eliminated or at least alleviated some of the problems. One of the test subjects, when asked why he/she did not use the help system, answered that it would probably not be needed. Another commented on how the button for activating it was hard to find and identify as it was so similar in appearance to the background.

Nielsen's heuristics are very clear regarding consistency. The fact that the button was not easy to find was perhaps due to not being found in the standard "Help" menu item or using one of the standard icons for online help. Instead, it was shown as a flat button somewhere in the top-right corner of the various frames. A standard icon was, however, used in the main window, adding to the general inconsistencies. The help button found in the data entry frames is shown in Figure 4.2.

The progress and status of the diagnosis is shown using a tree view widget, as shown in Figure 4.3. Upon starting the application, the tree is collapsed and it is unfolded gradually as the analysis takes place, highlighting the current state and the partial diagnosis via checking it off in the tree. Once the tree unfolded, it appeared to be clear to the majority of the test subjects what it did. The tree was in an editable state, and it was found to be confusing that editing it did not have any effect on the system.

The heuristics for designing intuitive graphical user interfaces do not provide a simple

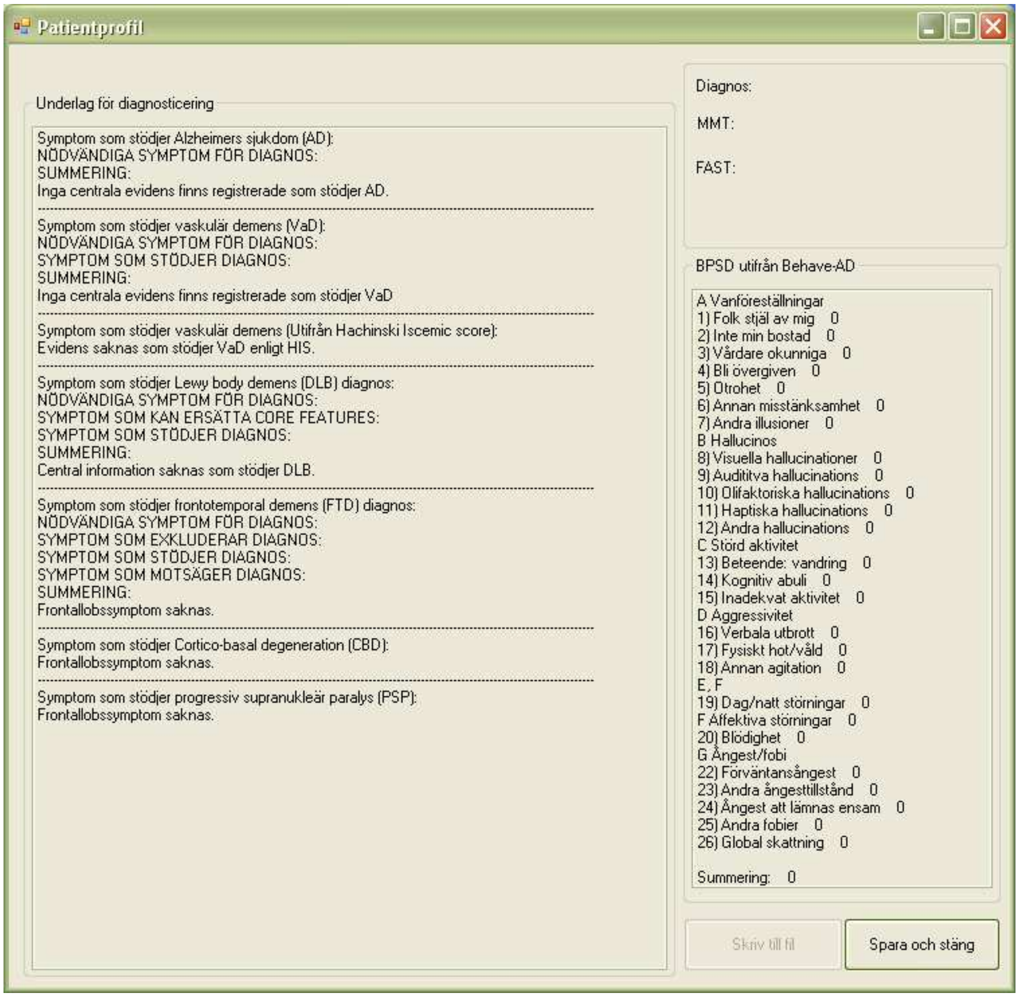


- ☑ **Investigation of dementia**
  - ☑ Existence of cognitive disease
    - ☑ Present
    - ☐ Absent
  - ☑ Type of cognitive disease
    - ☐ Amnestic syndrome
    - ☐ Delirium
    - ☐ Mild cognitive impairment (MCI)
    - ☐ Non-fluent aphasia
    - ☑ State of dementia
  - ☑ Type of dementia
    - ☑ Typical (one single unambiguous diagnosis)
      - ☑ Alzheimer's disease
      - ☐ Vascular dementia
      - ☐ Lewy body dementia
      - ☐ Frontotemporal dementia
    - ☐ Atypical (mixed or rare diagnoses)
    - ☐ Dementia due to other cause (ICD-10)
  - ☐ Intervention
  - ☐ Follow-up

Figure 4.3: Tree view widget showing the status of the diagnosis and progress of the analysis.

solution to the task that the tree view widget attempts to solve — displaying which of the “branches” in the decision tree that has been identified as the correct one. However, they do state that errors should be prevented if at all possible. Leaving the tree in a modifiable state when these modifications are not regarded by the system is clearly an oversight that only leads to confusion.

Once the analysis has finished, the data is presented in the “Profile” frame. A screen shot from the first evaluated version of DMSS-R of the frame in question is shown in Figure 4.4. As shown, the data is presented in a textual way and the sections are separated by a sequence of dashes. Also evident from Figure 4.4 is that non-relevant sections are shown, sections about types of dementia where no proof or any information is presented. The “Profile” frame is a problem from a learning perspective (see Section 4.3). From a user interface perspective, it was observed and expressed by test subjects that it did not provide a clear overview. Test subjects reported having difficulties “getting the message” that the contents of the window conveyed, and having to read the contents in its entirety. Patel et al. have showed that algorithmic representation in guidelines is to be preferred over textual [40]. Thus, perhaps the match between the system and the real world would improve if a similar form of representation was implemented in DMSS-R.

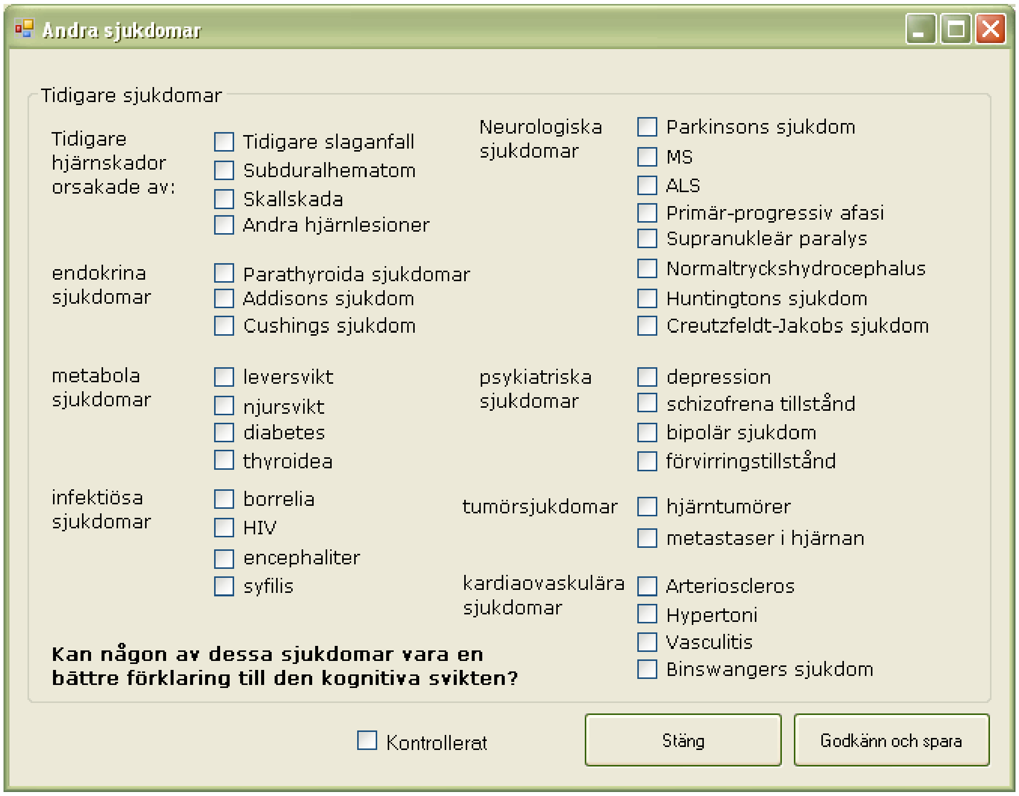


```
Symptom som stödjer Alzheimers sjukdom (AD):
NÖDVÄNDIGA SYMPTOM FÖR DIAGNOS:
SUMMERING:
Inga centrala evidens finns registrerade som stödjer AD.
------------------------------------------------------------
Symptom som stödjer vaskulär demens (VaD):
NÖDVÄNDIGA SYMPTOM FÖR DIAGNOS:
SYMPTOM SOM STÖDJER DIAGNOS:
SUMMERING:
Inga centrala evidens finns registrerade som stödjer VaD
------------------------------------------------------------
Symptom som stödjer vaskulär demens (Utifrån Hachinski Iscemic score):
Evidens saknas som stödjer VaD enligt HIS.
------------------------------------------------------------
Symptom som stödjer Lewy body demens (DLB) diagnos:
NÖDVÄNDIGA SYMPTOM FÖR DIAGNOS:
SYMPTOM SOM KAN ERSÄTTA CORE FEATURES:
SYMPTOM SOM STÖDJER DIAGNOS:
SUMMERING:
Central information saknas som stödjer DLB.
------------------------------------------------------------
Symptom som stödjer frontotemporal demens (FTD) diagnos:
NÖDVÄNDIGA SYMPTOM FÖR DIAGNOS:
SYMPTOM SOM EXKLUDERAR DIAGNOS:
SYMPTOM SOM STÖDJER DIAGNOS:
SYMPTOM SOM MOTSÄGER DIAGNOS:
SUMMERING:
Frontallobssymptom saknas.
------------------------------------------------------------
Symptom som stödjer Cortico-basal degeneration (CBD):
Frontallobssymptom saknas.
------------------------------------------------------------
Symptom som stödjer progressiv supranukleär paralys (PSP):
Frontallobssymptom saknas.
```

```
Diagnos:

MMT:

FAST:
```

```
A Vanföreställningar
1) Folk stjäl av mig    0
2) Inte min bostad    0
3) Vårdare okunniga    0
4) Bli övergiven    0
5) Otrohet    0
6) Annan misstänksamhet    0
7) Andra illusioner    0
B Hallucinos
8) Visuella hallucinationer    0
9) Audititva hallucinations    0
10) Olifaktoriska hallucinations    0
11) Haptiska hallucinations    0
12) Andra hallucinations    0
C Störd aktivitet
13) Beteende: vandring    0
14) Kognitiv abuli    0
15) Inadekvat aktivitet    0
D Aggressivitet
16) Verbala utbrott    0
17) Fysiskt hot/våld    0
18) Annan agitation    0
E, F
19) Dag/natt störningar    0
F Affektiva störningar    0
20) Blödighet    0
G Ångest/fobi
22) Förväntansångest    0
23) Andra ångesttillstånd    0
24) Ångest att lämnas ensam    0
25) Andra fobier    0
26) Global skattning    0

Summering:    0
```

Figure 4.4: The patient profile window frame from the first version of DMSS-R (scaled to fit page).



Andra sjukdomar

Tidigare sjukdomar

Tidigare hjärnskador orsakade av:
□ Tidigare slaganfall
□ Subduralhematom
□ Skallskada
□ Andra hjärnlesioner

endokrina sjukdomar
□ Parathyroida sjukdomar
□ Addisons sjukdom
□ Cushings sjukdom

metabola sjukdomar
□ leversvikt
□ njursvikt
□ diabetes
□ thyroidea

infektiösa sjukdomar
□ borrelia
□ HIV
□ encephaliter
□ syfilis

Neurologiska sjukdomar
□ Parkinsons sjukdom
□ MS
□ ALS
□ Primär-progressiv afasi
□ Supranukleär paralys
□ Normaltryckshydrocephalus
□ Huntingtons sjukdom
□ Creutzfeldt-Jakobs sjukdom

psykiatriska sjukdomar
□ depression
□ schizofrena tillstånd
□ bipolär sjukdom
□ förvirringstillstånd

tumörsjukdomar
□ hjärntumörer
□ metastaser i hjärnan

kardiaovaskulära sjukdomar
□ Arterioscleros
□ Hypertoni
□ Vasculitis
□ Binswangers sjukdom

Kan någon av dessa sjukdomar vara en bättre förklaring till den kognitiva svikten?

□ Kontrollerat

Stäng

Godkänn och spara

Figure 4.5: The window for entering previous diseases into the system (scaled to fit page).

Two data entry frames are treated differently than the others; the ones for entering previous diseases and laboratory results in the first two versions required the user to mark a check box with the word (Swedish "[k]ontrollerat") marked "[c]hecked" or "investigated" (depending on translation). Since these check boxes were not present in the other data entry windows, none of the test subjects understood what they were there for. From the point of view of the system, it is important to verify that the user has considered the contents of the frame and has entered as much information as possible. However, this is evidently not conveyed successfully using the chosen term. The check box is superfluous in itself — if the button for saving and exiting the frame has been activated, it should implicitly mean that the contents of the frame have been considered (just like for the other frames). For this reason, the checkbox was removed in the third version of the system.

In addition to the confusing check box, the window for entering previous diseases contains text shown in Figure 4.5 (text in Swedish, image from the second evaluated version of DMSS-R). The text is formulated in such a way that it is unclear whether *all* previous diseases should be checked off, or just the ones that might offer an alternative (better) explanation for the cognitive impairment. This ambiguity confused all of the test subjects, and the intention of the window should be presented clearer. The contents of the window in itself is a good learning and reasoning aid, because it acts as a checklist and reminder that some dementia-like symptoms may have other explanations than dementia.

4.3 Learning

Non-expert learning is a highly relevant part of the evaluation study, and one that has several aspects. Due to being non-experts, the test subjects are assumed to still be learning how to perform reasoning and investigation in the medical domain (psychogeriatrics in particular). We have already discussed the support for reasoning and medical investigation in Section 4.1 and the problems of learning the graphical user interface in Section 4.2. In this section, we focus on how a user learns how to use DMSS-R and how the medical learning process may be supported by DMSS-R.

Learning to use the system is an activity in itself. Using Activity Theory terms, its operations are common computer usage operations such as handling input devices and parsing screen output. Some of the actions are:

- to learn what the buttons in the main screen do;
- to understand how to start using the system;
- to learn how data is filled in;
- to learn how to obtain help and refresh medical terms that the user may have forgotten; and
- to learn how to request that the system suggests a diagnosis, to name but a few.

The list above is not exhaustive, as users have different experience and knowledge. Rather, the elements in the list are some of the core steps that can safely be assumed that all users will go through and consider to be actions at some point. Some users will quickly automatise the actions, whereas others have more difficulties in doing so.

4.3.1 Learning to use DMSS-R

The usability problems presented in Section 4.2 made it unnecessarily hard for the test subjects to learn how to use DMSS-R. Since the details of the problems have already been discussed, time shall not be devoted to them in this section as well. Instead, we note some of the conceptual difficulties with using DMSS-R:

- understanding how to begin using the system;
- entering data into the system; and
- understanding what state in the analysis process the system is in.

The first problem users experience when confronted with DMSS-R for the first time is that there seems to be no logical way to start using it. This has been shown in all sessions of the evaluation study, with the exception of the one with the expert who had already used the system in previous iterations (before this study). The intention from the designer of the main screen is that the user should read a text that describes the process. However, this was not the case. There may be several non-exclusive explanations to this:

- the text is too long, and therefore does not seem like a “quick help” text, thus not worthy of attention while trying to figure out how the system is used;

- the box in which the text is presented is not visible enough (being a shade of grey with black text); and/or

- due to the text only being presented as plain text (i.e., no images or different font sizes or text decorations), it does not grab the attention of the user.

Whatever the reason may be, it does not appear to help the users get started with the system. Naturally, once the users had begun using the system, they seemed to be perfectly fine with using it and could easily navigate between data entry windows. But, considering that a general practitioner may only meet as few as 1–4 new dementia patients per year [24], every use of the system might feel like a new beginning.

The specifics of the problems related to how to enter data into the system have already been covered at length in the previous section and will not be reiterated. The conceptual problem of entering data is that of test subjects' feeling that it was hard to state with certainty either that a problem was present or that it was not. The test subjects felt that the true answer may lie in between. The system, while allowing for entering severities in some cases, requires either a firm "yes" or "no".

That the test subjects found it hard to choose the correct answer on a scale that only allows for either indicating the presence or the absence of problems can be explained, at least in part, by the fact that they are non-experts. In contrast, the expert had no such problems answering questions of this type. This phenomenon is related to what was stated in Section 1.3 about differences between experts and non-experts. The expert could use experience, rules of thumb and a more well-structured knowledge base to make judgement calls quicker and with more confidence than the non-experts.

All test subjects clearly showed, and many expressed to have, problems determining in what stage of the analysis process the system was. This is a symptom of the test subjects having difficulties learning how to use the system, as it shows that they did not understand what the system did and how it performed its task. The first version required the user to manually advance the system through the process via pressing the button for analysis in the user interface. The lack of feedback combined with the lack of understanding of the given feedback (i.e. the diagnosis tree and text box) proved to be very troublesome and an unnecessary hurdle to overcome in learning how to use DMSS-R. The second and third versions did not require as many clicks as the first, but still required the user to click the "Analysis" button again once missing data had been entered.

4.3.2 Using DMSS-R to learn

DMSS-R is not aimed at teaching a layperson how to diagnose dementia; even the least experienced intended users are highly educated physicians. Its main responsibility is to provide these physicians with decision support. However, to successfully provide decision support to non-experts, the system must at the very least contain reminders to help the user fill in the correct values. Without correctly entered data, the system is unable to perform its task. In Section 4.1.3, we briefly discussed this issue in conjunction with the great amount of trust displayed by most of the test subjects toward the system.

DMSS-R offers online help on some medical terms via buttons in the interface. Only terms connected to a data input widget have help buttons attached to them, whereas other (for instance, in the window for entering lab results and previous diseases) terms had no such support in the evaluated versions. There were several problems with the information text itself:

- the text was not always informative (as an example, the text for Swedish "omdöme" — judgement — merely phrased the term as a question, asking if the capabilities were affected);
- nowhere were the sources of the definitions or descriptions to be found; and
- no reference was provided to further suggested reading.

These problems make it very hard for the non-expert to find more information on a topic. van Merriënboer et al. discuss how learners require integrated support systems and prefer them to external ones [49]. As an analogy, they compare training wheels on bicycles (integrated support) to a parent running beside the child learning to ride, yelling to constantly remind the child to keep the steering wheel in an appropriate position (external support). The ill effects of relying on external support is that it overloads the cognitive resources of the learner by having to switch contexts often and thus adding the old context to an already strained memory. DMSS-R should offer better support in this regard, so that non-expert users do not need to have other sources of information handy to successfully use the system.

The help for understanding what the levels of severity mean was, during the first two versions, only present in the general help section of the program. On the screens where the severity input widgets were present, there was none. Obviously 2/2 is worse than 1/2, but without a frame of reference, it is not clear what these numbers mean. Again, the system did neither provide easily located sources nor references to other suggested readings that would help the user learn about the levels and how to correctly gauge severities. The set of rules used for inference could be located via buttons in the main window, but no references were made to how to assess the level of severity were made explicitly. The third version made the levels more explicitly clear by replacing the numbers with text labels, and by adding a text frame where the level was explained in terms of whether the symptom was so severe that it did not (mild) or did (severe) impact ability to work and function normally. The addition of the text frame was observed to be helpful, once the test subject had noticed that the text existed.

Section 4.2.2 discussed the user interface problems related to feedback. One of these problems is a particular problem from a learning perspective, namely the lack of explanation *why* a certain missing data item has been marked with red colour (other than conveying the message that the item is required in some way). Moreno has studied the effects of corrective (being told that one is wrong) versus explanatory (being told why one is wrong) feedback, and found that learning increased significantly when students were given explanatory feedback [32]. Regrettably, current versions of DMSS-R provide only corrective feedback, and is therefore not as efficient as a learning tool as it could be.

4.4 Attitudes toward DMSS

As stated in Section 3.2, the questionnaire from Appendix B was used to investigate the attitudes of the test subjects toward using DMSS. The questionnaire was given to the medical students and the intern, but not to the expert who is part of the DMSS development team (for bias reasons). The full set of answers is presented in Appendix C.

The results show that the test subjects are overall positive toward the system, but that it did not affect what investigation steps or treatments would be recommended for the patient. 7 of 9 stated that they agreed to a large extent (or more) to the statement that the system may act as an important support for their future activities as physicians. All stated the same agreement to the statment of being positive toward the system. On the negative side, of the majority of 5–6 who stated an opinion on being affected by the system regarding what investigation steps and/or treatments that would be recommended for the patient, the attitude was more skewed toward only agreeing to some extent. Both the positive and negative results should be interpreted bearing in mind that three of the test subjects were affected by the system bug that prevented them from obtaining a diagnosis. The ones who did not obtain a diagnosis were understandably generally more negative toward being affected by the diagnosis suggested by the system (because there was none), whereas the more positive answers came from test subjects who obtained one.

As for ease of use, the attitude was largely positive. 2 of 6 agreed to some extent with the statement that the system was easy to use, 1 agreed completely and the remaining 6 agreed to large extent. Similar positive responses indicate that the test subjects would like to use the system (or similar) and recommend it to their colleagues.

5 Suggestions for improvements

In this chapter, suggestions for future version of DMSS-R are presented, based on the results of the evaluation study and on the theoretical foundation presented in previous chapters. The suggestions should serve as topics for further evaluation studies, and may act as milestones in the iterative development of upcoming versions of the software system. Some suggestions have already been incorporated to some extent in development versions of DMSS-R, whereas others are yet to be implemented.

5.1 Information and help frame

In the first two versions of DMSS-R that were used during the study, information and help about the various data entries could be obtained via clicking on a button, which resulted in displaying a standard Windows message box[1] containing the text. This is a troublesome design, since it requires much interaction from the user, and because the text must disappear for the user to be able to make the entry comfortably. Thus, it requires the user, who is obviously struggling with remembering the definition of a medical concept (otherwise he/she would not have requested the information in the first place), to remember the definition or explanation in addition to whatever the correct entry should be — at the same time as the user has to manipulate the interface by closing the message box.

Choosing the message box as the medium for displaying the information also poses some limitations (some of which are technical) on the contents. Most importantly, it limits the amount of text that can reasonably be presented, in particular since standard message boxes do not allow for scrolling functionality. Also, images are not easily presented in text message boxes, and adjusting the fonts to allow for headings is also troublesome.

As a solution, we suggest that a portion of the data entry frame(s) is dedicated to information and help. The widget chosen for this task should be able to display text in various fonts and sizes, images, and hyperlinks that can open either web sites or further "rich" dialogue boxes.[2] The widget must also be scrollable. The contents of the widget should also be improved and extended to include the following:

- a header, clearly stating what the information pertains to;
- the medical definition (stating the source of the definition and providing a hyperlink to it) of the term in question;
- text describing how to make the judgement between the various degrees of affection, with hyperlinks to the guidelines that govern these degrees;
- a descriptive text, which may be easier to follow for a non-expert than the medical definition; and

[1] Consisting of a text message and a button labelled "OK".

[2] Ones that can display the aforementioned elements.

– a text describing, in general terms, the importance of the concept in relation to dementia diseases.

The contents of this information and help frame should be changed via a click on a button, like in present versions of DMSS-R. When the data entry frame is first displayed, the information and help portion should display a message stating that clicking on the button next to the medical term will display the information in the frame. Since the contents of the frame changes only when the user clicks on the information request button, a header is required as a means of clearly conveying what the information in the frame pertains to.

For the non-expert, it has been shown (both in this evaluation study and in the literature [40, 8]), that uncertainty is an impeding factor to making decisions. Deciding whether a person has an affected semantic memory, or an unaffected one, is hard to do without rules of thumb or clear support from guidelines. Thus, in order to support this important process and get correct data for DMSS-R to work with, there is a need for the correct definition as well as the relevant guidelines to be embedded in the system.

Given that one of DMSS-R's most important market segments is made up of the general practitioners that do not come in contact with many dementia cases, refreshing their memory on the theoretical background and why certain concepts are important to diagnosing dementia is a feature we deem useful. Even more so since medical knowledge increases rapidly, and it is infeasible for a physician to keep updated on all recent developments.

The third version of DMSS-R has adopted this suggestion to some extent. The interface has a reserved text box for information and help, but the text is still limited to short messages and it is still presented as plain text. References to the information source were added, but without hyperlinks to the original source.

5.2 Data input feedback widget

A problem, as stated by one of the test subjects directly but observed among all of them, is that of knowing exactly what has been entered into the system. This is true in particular when the data input widget for entering severities was used. To remedy this situation, we suggest that in addition to the information and help frame that shall occupy a portion of the data entry frame an additional frame occupies another portion. That portion would house the data input feedback widget, and it should display the data as it has been entered in a textual form. The entries in this widget should be made in a stack-like fashion (last in, first out), making the most recent entries the topmost ones. The widget should not merely record a history of all entries made, it should allow for at most one entry in the widget per term in the data entry window. Thus, if a change is made (for instance, changing from unaffected to affected), the entry in the data input feedback widget concerning the term that was subject to the change is removed, and the new entry from the same term is placed in the topmost position of the widget.

The text in the widget should use emphasis to clearly state what value has been entered, for example like below:

> The patient **has impaired semantic memory**, as defined in XYZ.

The text should be unambiguously formulated, to reduce the likelihood of misinterpretation. Severities should also be clearly stated in a term-specific way (similarly to the information and help frame from the previous section), such as:

> The patient suffers from **mild impairment** of **semantic memory**. Severity **1/2**, defined in XYZ.

Where XYZ in the examples is a hyperlink to the guideline and its relevant section that dictates and defines the levels of severity. The term "mild impairment" must coincide with the terms used in the guidelines.

Combined with the information and help frame, the goal is to reduce the amount of uncertainty of *how* to enter correct data, but also to show *what* data have been entered into the system. Thus, the overall cognitive load and stress from uncertainty is reduced and the hyperlinks to relevant sources of information help in directed learning efforts by the user.

The third version of DMSS-R did not have two separate text boxes, it featured a single box used for both help and for displaying the latest entry that had been made. Once test subjects had noticed this, it appeared to help them understand what data they had just entered into the system. However, the text was plain text and did not feature the bold parts nor did it have any hyperlinks to further reading.

5.3 Updated data input widget for severities

The changes made to the data input widget in the three versions of the system were successful in removing most of the complications related to misinterpreting the colours. However, no satisfactory progress has been made to make entering severities intuitive. With the current scheme, the users simply do not understand how (or even *that*) they are supposed to perform the task.

At the heart of the problem may lie that the widget for entering severities, and the regular data input widget are too similar. Due to similar appearance, the user is fooled into thinking that the widgets behave in a similar way. To combat this problem, we suggest that the widget for entering severities is changed and made dissimilar to the regular data input widget. This will more clearly show that they are used in different ways.

In the third version, the severity input widget was changed to look as in the bottom of Figure 4.1. As shown, the level of severity is displayed as text in the box, rather than as numbers. Clicking the button for presence of abnormal features a single time displayed, in Swedish, "grad?" ("severity?"). This indicates that it has been established that there is a problem, but that the severity of the problem is unknown. Subsequent clicks on the button cycles the levels of severity in increasing order, eventually resetting and coming back to "grad?". A test subject using the third version of the system commented that it was nice to be able to state the presence of a symptom without having to specify the severity of the problem.

Based on the findings of the evaluation study, we suggest that other types of severity input widgets should be tested and evaluated. In designing the widgets, the guiding principles and parameters are:

- no matter the number of options or levels, the amount of screen real estate must be constant (during active use of the widget, the amount may be increased);
- the widget should be obviously dissimilar to the regular data input widget; and
- it should be easy to select a single, correct value using both mouse and keyboard.

Some suggestions for these widgets are presented in Figures 5.1, 5.2, and 5.3. Figure 5.1 shows a single drop-down menu approach. It is graphically and conceptually different from the data input widget, which ensures that the users quickly understand

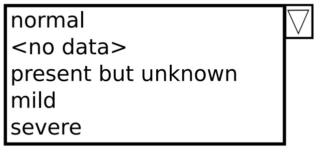


Figure 5.1: Suggestion for severity input widget using a single drop-down menu.

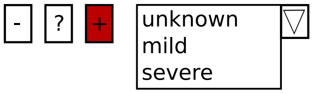


Figure 5.2: Suggestion for severity input widget using both familiar data input buttons and a drop-down menu.

that the kind of data it can capture is not the same as the ordinary input widget. While this clear break may be beneficial, a problem is finding the appropriate phrases to use in the drop-down menu. Also, previous versions of DMSS-R used drop-down menus extensively and they were showed to present usability problems — it was too easy to enter the wrong data by accident.

A middle-ground between the current system of buttons and the drop-down menu is shown in Figure 5.2. Initially, the drop-down menu is in a disabled state. The menu is disabled until the "+" is activated, to clearly show that the data one can enter via the drop-down menu is only used in presence of abnormalities. In addition to being disabled, the contents of the box housing the drop-down menu is the empty string, to avoid confusion in that (for instance) the green button has been activated and yet the text says "mild". As shown in the figure, the amount of choices in the menu is reduced due to the buttons having semantic meaning as well.

Using the concept of combining the buttons and some other widget to enter severities, Figure 5.3 shows a suggestion using the buttons and a slider. Similarly to the suggestion with buttons and drop-down menu, the slider would also be in a disabled state until the "+" is pressed. The slider clearly presents the possible values to the user, which is intended to promote understanding that the level must be entered as well.

Both suggestions using the buttons in addition to some other widget are at risk of being used similarly to the current severity widgets. However, both suggestions should be more intuitive than the current system since both drop-down menus and sliders are commonly used in user interfaces, whereas buttons that require repeated presses are not. These suggestions should be evaluated in a future study.

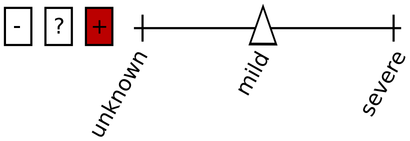


Figure 5.3: Suggestion for severity input widget using both familiar data input buttons and a slider.

5.4 Clearer and explanatory feedback for missing data

When DMSS-R has determined that data is missing, it should notify the user of this in a clear way. The first two evaluated versions of DMSS-R were suffering from some inconsistencies in this area. We suggest the following:

– if a data entry point requires an entry, the text for the entry should always be made red *and* boldfaced; and

– a clickable button with an icon should appear to the far left.[3]

The first suggestion is due to the current inconsistent behaviour of the window (see Section 4.2.2) of sometimes not marking the entries themselves, but rather the frame surrounding them. Additionally, increasing the level of dissimilarity by making the text boldfaced makes the text easier to locate for colour-blind users.

The click-able button also serves the purpose of quickly displaying what entries are required, but most importantly, it serves as a mean to allow the user to access explanatory feedback. When the button is pushed, the user should be presented with a clear explanation of why the item has been marked. It has been shown that novices learn much quicker if they are given explanatory rather than corrective feedback [32]. That is, if they can somehow learn *why* something is wrong rather than just being told that it is. Present versions of DMSS-R do the latter, but to act as a tool that mediates learning, it must do the former. To do so in a clear and easily computer-generated way, we suggest a graphical display of the logic that drive the reasoning process. It is presented in the next section.

van Merriënboer makes a distinction between supportive information and procedural information [49]. The former is concerned with support for problem solving, whereas the latter is related to how tools are used. To increase the effects of supportive information, a system should provide elaborate explanations. These elaboration techniques help the learner to formulate schemas, into which information and knowledge are coded and made available for later use. In doing so, the gap between the current knowledge of the learner and the material that is to be learnt is bridged. This is the kind of support that is needed for DMSS-R to truly become a valuable tool for learning.

5.5 Graphical display of reasoning logic

As previously stated, learners are supported far more if they are given explanatory rather than merely corrective feedback [32]. Also, it has been shown that the current patient profile window in DMSS-R offers little overview, forcing the user to scan a large corpus of text. In doing so, the user must construct a model of the reasoning that DMSS-R has undergone, and verify its correctness — even more so if the system could not conclusively infer a suggestion for diagnosis. The text is also presented in an order not related to the actual findings of the analysis process, but to how common the various types of dementia are, presenting the most common (Alzheimer's disease) at the top. Even types of dementia for which there have been no evidence at all are presented in the text, adding clutter and unnecessary cognitive load.

We suggest that the reasoning process should be presented to the user in a graphical way, rather than textual. For this, we suggest that the system uses an interface similar to many graphical logic systems, where nodes represent evidence in the reasoning process and the user can follow all inferences made. One such logic system is Araucaria,

[3] The text is aligned to the right, thus leaving the left side of the text free and therefore making an icon appearing in that space highly visible.

developed at the University of Dundee. While it has not been proved conclusively that graphical systems improve the reasoning skills of students, strong tendencies seem to point in that direction [48]. The suggested system offers several important features:

- non-experts are helped to learn and internalise the medical guidelines (hyperlinks should be provided to the relevant sections of the guidelines, as well);
- experts can verify the correctness of the system, thereby decreasing sceptical attitudes toward it (see [40, 46]);
- the graphical representation aids in understanding the causality between symptoms and diseases;
- missing or conflicting information becomes easier to identify, since the user does not have to compare two lists (one listing the requirements, the other listing the presently entered data) but rather gets an overview conveying the relation; and
- compared to the system in current use, only relevant information needs to be displayed, lowering the total amount of information presented in one screen.

Due to the reliance on *core features* in DMSS-R [24], and the fact that not all entered data necessarily are required by the medical guidelines, the total number of nodes that need to be presented on screen should not nearly equal the total number of possible data entries. For learning purposes, however, it would be useful if all nodes — both those that were relevant for inferring the disease and those that were not, could optionally be displayed.

The display should be clear, and contain information on what guidelines were used to guide the reasoning process. Access to these guidelines should be made available in the window via hyperlinks so that the user can acquire additional information.

Conceptually, the display might look as in Figure 5.4. Note that the sketch is not intended as a final prototype, nor has it been evaluated with users. It does, however, have the key elements that are required to fulfil the above list of intended features. As shown, conditions are displayed using ellipses. The topmost ellipsis is the condition in which the user is interested. The colours of, and the connectors to, the other ellipses are relative to this condition of interest. Required conditions are drawn using a single line until it has been proved whether they are present or absent. If a required condition is present, it is drawn with two lines and green colour. Provably absent conditions are drawn using a line that has been "crossed out" by two short lines, and using the colour red. As shown, colours are used to convey information, but the same information is available to colour-blind people using the line styles. The evidence required for disproving a condition is also marked in red, due to its relation to the condition of interest. Since there is contradictory evidence, the condition of interest in the upper part of Figure 5.4 is marked with red. Additionally, the topmost ellipsis is given a frame to differentiate it from a proved condition (which does not have a frame). In the lower part the figure, a condition with full support is shown. Both required conditions are evidently present, and as such, the condition itself is proved to be present.

We suggest that the graphical display should be used not only in the patient profile window, but in several other contexts as well, described in the upcoming sections. Using the same window consistently helps the users to internalise the guidelines and become familiar with the window itself. The new system must be evaluated carefully in further studies and further similar studies (like [48]) should be taken into consideration.

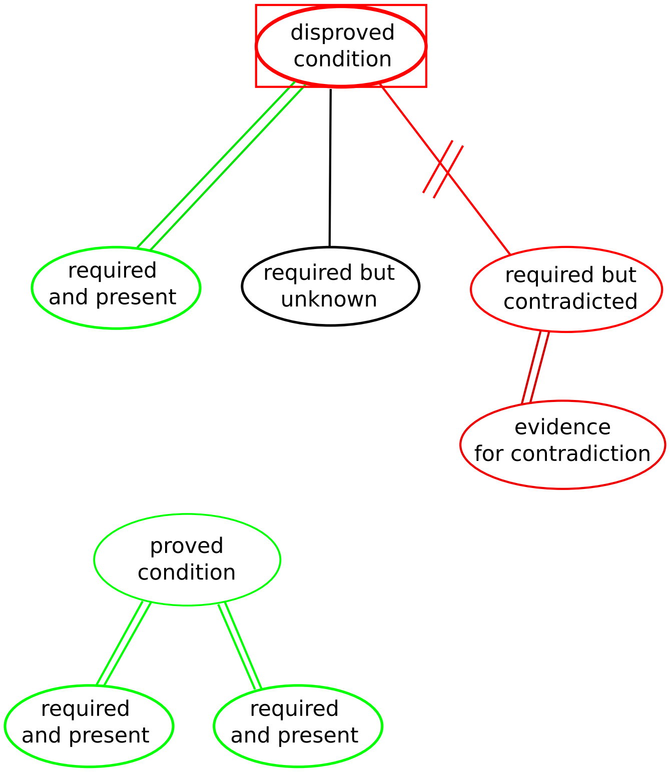


Figure 5.4: The suggested graphical reasoning display.

5.6 Support for planning next visit

A problem in Sweden today is that the queues for obtaining medical care are increasing, and reducing them so people are treated in a timely fashion has become an important political issue. DMSS-R is currently able to to support decision-making, *given the appropriate data*. The system will alert the user if it determines that some additional entries are required, but it only does so in a reactive manner — the system cannot currently inform the user that some entry *may* be needed, depending on other (missing) values.

We suggest that the system should try to, given the amount of data already entered, make an outline of tests that may be relevant during a future patient visit. This could theoretically reduce the amount of visits that are needed. The evaluation study and results from the literature agree that non-experts are insecure and requisition more tests to be made than an expert. Extending the capabilities of DMSS-R to guide the decision of which tests to make during an upcoming visit might help in reducing the amount tests that are made, thus lowering the total costs and the time effort required for diagnosis. The system should, of course, clearly state the reasons why some tests should be made in order to aid learning.

5.7 Adjustable help and information system

The help and feedback systems presented in previous sections are all designed to be useful, but the are geared toward the non-expert user. Research shows that unnecessary information increases the cognitive load [49, 39], even for users who try to ignore it because they know it by heart (the act of actively ignoring it requires cognitive resources). Furthermore, it has been shown that support for various levels of expertise must be adaptable to suit its intended audience [49, 40]. Thus, it should be possible to choose the level of support one requires from DMSS-R, ranging from complete novice to expert. The expert might, for instance, find the more casual description of terms and concepts in the information and help frame useless. It should therefore be possible to disable various parts of the help and information system.

In addition to being able to disable certain parts, it should also be possible to set which sources should be used for obtaining definitions and other information. This makes it possible for medical institutions to choose what set of guidelines should be used, for whatever reason. A planned feature for DMSS-R is to allow the user to select the set of medical guidelines that are used during analysis. Thus, it would be sensible to show the effects of this choice in the user interface as well.

5.8 Replacement for decision tree widget

The three evaluated versions of DMSS-R have a tree widget for displaying both the logically inferred diagnosis and the status of the analysis progress as a whole. It does so by displaying the tree incrementally, opening branches only if one of the sub branches has a check mark in them. The tree does not, however, show *how* the inference has been made, nor what is missing or conflicting that prevents it from advancing in the analysis.

Using the graphical display of reasoning logic that was suggested in Section 5.5, these shortcomings are remedied. It would provide the additional information that would help the user understand what “goes on” in the system, without having to worry about understanding the logic behind it.

One of the design reasons for choosing the tree widget was to give the user an overview of the process. The process, however, is in the current version of DMSS-R at least one and at most three steps long (not including steps beyond getting a diagnosis). Additionally, many of the branches are mutually exclusive — a patient cannot both suffer and *not* suffer from dementia at the same time. We therefore suggest that the widget is replaced by a progress bar and a bulleted list of statements that the system can make by inference. The progress bar would be clearly labelled as an indicator of the analysis progress as a whole and the bulleted list would state each fact and include a hyperlink to a graphical representation displaying the reasoning that has proved the fact. Something similar to the following suggestion:

– The patient suffers from **XYZ** (show reason)

In the example, “show reason” is a hyperlink that opens a graphical reasoning display window.

If, for some reason (i.e. conflicting evidence) DMSS-R cannot determine what the patient is suffering from, the progress bar will not go to its full position. It should stop at an appropriate point, depending on how far the system could get in the process, and there should be a text describing the situation. For example:

> The system could not continue further, due to conflicting evidence. Click here to see the reasoning process, and why the system could not continue.

The information presented in the current version of DMSS-R via the tree widget is limited to only a few points. Presenting these in a tree shows the user what other possibilities the system has considered, but makes it hard for the user to get an overview — and requires scanning the entire structure to find the marked entries. Presenting the results in a bulleted list gives a greater overview (considering that there will be at most approximately three items in the list, the list itself will be very short and require less time to scan than the tree structure), and the links to the graphical display of the automatic reasoning process provide the non-expert with valuable information useful for learning and for refining their internal reasoning processes.

5.9 Adding a tutorial that also acts as a learning aid

DMSS-R does not currently come with any help or documentation, other than that which can be obtained via the help button. As the evaluation study has shown, users do not activate this feature, and therefore do not benefit from it. While this creates problems related to incorrect use of the program, one troublesome aspect of the current system is that users must use data from their actual patients, trusting the system before it has earned their trust and confidence.

We suggest that the users should be offered a tutorial, which could serve both the purpose of learning how to use the operations of the system as well as learning how to perform the activity of using the system to obtain a diagnosis. The tutorial would be an add-on to the program, since it would be tailored not just toward teaching how to enter data, but also act as a computer-based tutoring system for medical problem-based learning. One such system is called COMET — however, it also focuses on collaboration between problem-solvers [45]. Using COMET or a similar software as a component, DMSS-R could become a learning aid and an integrated part of how medical students learn to diagnose dementia. It has been shown that the use of COMET has improved the clinical reasoning skills of students significantly, compared to those being tutored by a human [45].

5.10 Enforcing consistency and increasing clarity

Section 4.2 stated several problems relating to poor consistency and lack of clarity. These problems should be considered errors and be taken care of. Special care should be taken to ensure that the resulting system is accessible to colour-blind users. This includes not relying on colour to be the only differentiating factor that highlights important things that require the attention of the user. Icons should be used to greater extent (cf. Section 5.4) to aid in this, as should font changes (e.g. boldfaced font). Tooltips should be provided to inform the user what the various buttons do.

Most importantly, all text should be written in a clear way, so that ambiguities are avoided and instructive text clearly states its intention. In particular the text in the main window of DMSS-R should be rewritten so that users can more easily understand what to do and what the system requires of them. To this end, it is imperative that the widgets for showing text are capable of handling headings, emphasis, and so forth.

6 Discussion

This chapter is devoted to discussion of some of the topics that have impacted the results of the study. First, Section 6.1 discusses the attitudes toward the system. The general attitudes were found to be positive, which is a surprising result compared to other similar studies. In Section 6.2, we discuss one of the biggest problems of the study, namely the scarcity of test subjects. Less than a quarter of all medical students that were interested in participating in the study actually did so, and none of the interns at the psychogeriatrics ward participated. We try to identify some of the causes of this great absence so that future studies can be more successful. Apart from the low number of participants, a large problem was that of time. It is discussed in Section 6.3, in addition to discussing the methods used during the evaluation. Finally, the system bugs related to obtaining a diagnosis are discussed in Section 6.4. Studies with prototypes are often subject to bugs, much to the dismay of both test subjects and the evaluators.

6.1 Attitudes toward the system

The study shows that DMSS-R has several problems which prevent its usefulness as a tool for mediating medical learning and reasoning in non-experts. Shortcomings in the graphical user interface have proved to be issues that cause breakdowns, and programming errors during the study (causing the system to be unable to finish its analysis) made the system hard to use. And yet, in spite of these factors, the questionnaire for investigating the attitudes toward the system shows positive results.

How may this positive attitude toward the system in general be explained? Why are the participants in this study more positive to this system than the physicians in the study by Toth-Pal [46]? The test subjects were younger, and perhaps more inclined to viewing the computer as a source of information rather than problems. Perhaps it may be explained by the fact that the system "knows" more than they do, and they feel as if they may learn from the experience of using the system.

Other studies show that experts tend to view medical guidelines in a positive light, although they do not use them to large extent [40]. Non-experts, on the other hand, find it hard to relate the ideal case presented in the guidelines to the patient case they are currently dealing with. This difficulty may offer an explanation to why the non-experts did not feel that the system supported their reasoning. Their positive attitude contradicts what Patel et al. found, because non-experts tend to find guideline-based systems hard to use for this very reason.

As stated previously, the study by Toth-Pal concerned a different type of system, one which was intended to be used while meeting patients. One of the main complaints against the system in that study was that it was distracting to use during these appointments. This might also be an important factor to the difference in attitudes toward the two systems. More work should be done to determine if the attitude toward DMSS-R is as positive as the results of the questionnaire study seems to indicate.

6.2 Scarcity of test subjects

About 40 medical students stated that they were interested in being involved with the evaluation study (almost all students attending the lectures). In the end, only eight — less than a quarter, of these actually took part. There are several plausible reasons for this large difference:

- the psychogeriatrics ward was stricken with the calici virus during one of the evaluation periods;
- the medical students spent only a single day at the psychogeriatrics ward, and many may therefore have felt ill-prepared for participating in the evaluation; and
- the hectic schedule of medical students may have prevented many from participating in the study (each session took between 30 minutes and one hour), in spite of previous interest in doing so.

Whatever the reason may be for the individuals who did not participate, it is clear that if a similar study should be conducted using medical students, these factors must be taken into consideration. As a consequence of the low number of participants, the study could not be made as a regular quantitative usability study. Also, the data is not large enough to draw too many general conclusions outside the population of test subjects. At best, the conclusions are indications that require further studies.

The geriatrics ward employed several interns during the semester. These were contacted, but none volunteered for the study. Had they participated, not only would the study have had more data, but it would have been possible to observe the learning process more closely and over a larger time span. The test subjects in the study, with the exception of the expert, only used the system during a single session. Interns could have used the system repeatedly, and would therefore not get caught up in mistakes common to all beginners during their first session with the system.

6.3 Methods used and time constraints

Due to the limited time during the sessions themselves being a factor, it would be interesting to consider how (1) more time during sessions; (2) introducing the system via mandatory reading of the help window and time to explore the system freely; and (3) more actual patient cases would impact the results of the study.

Using more time per session would have been theoretically possible, but considering the difficulties of getting test subjects to volunteer for sessions lasting between a half and one and a half hours, it is unlikely that longer sessions would have been more successful in bringing in test subjects. Increasing the incentive to participate by offering more than a lunch sandwich might have enabled longer sessions to take place. If the study should be repeated with medical students, it would be very beneficial to make participation in the study a requirement to pass the tenth semester (but it is unlikely that this is a possibility).

Lowering the threshold to begin using the system efficiently by mandating that all test subjects should peruse the help text might have sped up the sessions, allowing for more time to assimilate knowledge and therefore more data with regard to medical reasoning and learning processes. However, doing so could possibly have negatively impacted the amount of data with regard to usability concerns. Since no previous studies have been made with DMSS-R in that regard, it was deemed valuable to place the intuitiveness and ease-of-use under scrutiny.

It was clear that actual patient cases were faster for the test subjects to use, as opposed to the fabricated one. Therefore, increasing the amount of actual patient cases — perhaps not using the fabricated test case at all — could have yielded more results. However, using the set of test subjects that was available, it would not have been possible to increase the amount of actual test cases, due to the test subjects' lack of experience. Since the topic of the thesis was to study non-experts, the lack of experience is an inherent factor. Had the time frame been longer, it would have been possible to study general practitioners, rather than medical students. General practitioners are also non-experts, but have more experience with actual patients. Time and luck permitting, it would then also be possible to study developments in individual patient cases. Given the circumstances, we argue that the study in this thesis was carried out in the most sound way possible.

In particular the interview part of the evaluation sessions warrant some criticism. Since the study placed such importance on identifying breakdowns, it is regrettable if the interview questions were causes of such breakdowns. The purpose of the questions, as stated in Chapter 3, was to make the test subjects learn the graphical user interface of the system better and faster. However, the breakdowns may have lead to insecurity and negative conceptualisation. While conceptualisation helps learning in the long term, it may be negative in the short term sense. When the test subject began to work with the system again after the interview, the possible breakdowns caused by the questions may, in turn, have resulted in a large amount of cognitive resources being spent trying to relearn how to use the interface. Thus, the intended change in the object of the activity (the difference between Figure 3.4 and Figure 3.5) may have been negatively impacted even further, already severely affected by lack of time.

6.4 System bugs

As shown in the table with the diagnoses obtained by the test subjects from the system (Table 4.2), only one of the seven test subjects reached the correct diagnosis. It was believed that the second version of the system introduced a bug related to diagnosing Alzheimer's disease (due to underlying changes in the logic system). However, the table shows that even the first version had *some* bug, preventing it from reaching a diagnosis. Thus, both of the first versions had a bug (not necessarily the same) that affected the study.

The two diagnoses of "Mild cognitive impairment" (MCI) may stem from an inability of the test subjects to correctly gauge the severity of the cognitive impairments. As shown in a session with the expert, the difference between MCI and Alzheimer's disease may lie in the severity of symptoms — which is not very surprising, given that MCI may progress into dementia and Alzheimer's disease is the most common type of dementia. But, since both diagnoses of MCI were obtained from the third version of the system, it may be indicative of a bug introduced in that version of the system.

Whatever the reason may be, it is regrettable that only four of seven test subjects obtained a diagnosis in the fabricated test cases. Because of this, it was not possible to test how the other half of the test subjects verified their diagnosis. They were instructed to look at the patient profile window, to get some clue of how the system worked and where it had started having problems, but due to the profile window being unclear, they did not obtain much help in this way.

7 Conclusions

DMSS-R has been evaluated from three main perspectives: (1) as a tool for mediating medical reasoning and investigation for non-experts; (2) from a usability perspective; and (3) as a tool for non-expert learning (and as a tool to be learnt). Several problems with regard to these perspectives have been identified, and suggestions for improving the system have been made. Due to the limited amount of test subjects and actual patient cases, the limited amount of time during sessions, the results presented in this thesis can only be considered to be indicative. However, the results are valuable as such, and have lead and will continue to lead to significant improvements in the system. Some of the suggestions made in this thesis, mostly related to usability, were implemented in the third version of DMSS-R, dated 7th May 2008. The changes have been studied, and seem to have positive effects. As more of the suggestions are implemented, the system must be evaluated further to verify these preliminary positive findings.

The largest problem from the reasoning and learning perspectives has been identified as the lack of explanatory feedback. Without it, users cannot verify the reasoning/inference steps of the system, nor can they learn from it and internalise the reasoning process. First and foremost, a graphical display of the relationship between the evidence as data and the resulting diagnosis is believed to be beneficial as support for both reasoning and learning. Additional suggested features to improve the system in these aspects include:

- an information and help frame, displaying detailed information about medical terms;
- support for planning upcoming appointments;
- options for adjusting the amount of help and information displayed on screen (to avoid cluttering up the interface for experts who do not need the amount of help non-experts require); and
- a tutorial for learning how to use the system.

From the usability perspective, the biggest problems were related to data entry and feedback. Suggestions for taking care of these problems are:

- a data input feedback frame, where the entered data is shown and written in a clear way to avoid confusion;
- several suggested new widgets for entering levels of severity of symptoms;
- improved feedback for showing missing entries;
- a replacement for the decision tree widget that shows the status and progress of the analysis; and
- general consistency and clarity improvements.

Parts of these suggestions have been implemented, and as development continues these proposed changes will be evaluated during upcoming studies.

7.1 Future work

DMSS-R is constantly evolving, and both the suggestions for further development presented in this thesis as well as any other upcoming changes must be evaluated properly. These studies should ideally be performed using actual general practitioners working with real patients, and for a longer period of time. The time aspect is key, both due to the low influx of dementia patients and due to the learning process being time-consuming. Concurrently, the attitudes toward the system should be studied, and the positive results presented in this thesis should be verified, ideally with general practitioners of various levels of expertise rather than inexperienced medical students.

Due to the complications that this thesis has suffered, it is inadvisable that future work is conducted as a Master of Science thesis. The project should take at least one year to ensure that the learning processes can be studied, and because of this, would be better suited for a PhD student or researcher.

8 Acknowledgements

I would like to extend my sincere gratitude toward my supervisor, Helena Lindgren at the Department of Computing Science, Umeå University. She has helped me through the rough times when we did not manage to get any physicians to participate, given me interesting and inspirational articles to read, and implemented changes to DMSS-R, based on preliminary results of the work presented in the thesis. Her help has been invaluable.

This thesis would not be possible without the help from loved ones. Special thanks to my dear sister Lisa Larsson for proofreading the thesis on short notice, my wonderful friend Johanna Högberg for listening to me and supporting me throughout the whole process and, last but certainly not least, my bride-to-be Anna Nordström for all her loving support and patience while I worked on the thesis. I would still be stuck on the first page without all of you.

References

[1] A. Aaltonen. Eye tracking in usability testing: Is it worthwhile. *CHI'99 Workshop The Hunt for Usability: Tracking Eye Movements*, 1999.

[2] D. Avison, F. Lau, M. Myers, and P.A. Nielsen. Action Research. *Communications of the ACM*, 42(1):94–97, 1999.

[3] G. Bucht. Hjärnans åldrande. Lecture notes from Medical training at Umeå University, 2008.

[4] S. Bødker. Applying Activity Theory to Video Analysis: How to Make Sense of Video Data in Human–Computer Interaction. *Context and consciously: Activity theory and human–computer interaction*, pages 147–174, 1996.

[5] J.M. Carroll. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. Morgan Kaufmann, 2003.

[6] Apple Corporation. Apple Human Interface Guidelines. Online, last accessed June 13, 2008, 2008. `http://developer.apple.com/documentation/UserExperience/Conceptual/OSXHIGuidelines/`.

[7] Microsoft Corporation. Windows Vista User Experience Guidelines. Online, last accessed June 13, 2008, 2008. `http://msdn2.microsoft.com/en-us/library/aa511258.aspx`.

[8] B.J. Daley. Novice to Expert: An Exploration of How Professionals Learn. *Adult Education Quarterly*, 49(4):133, 1999.

[9] P. Eklund, R. Helgesson, and H. Lindgren. Towards Refinement of Evidence Using General Logics. To appear in proc. ICAISC2008, 2008.

[10] Y. Engeström. Expansive Learning at Work: toward an activity theoretical reconceptualization. *Journal of Education and Work*, 14(1):133–156, 2001.

[11] S. Eriksson. Lecture. Lecture notes from Medical training at Umeå University, 2008.

[12] GNOME. GNOME Human Interface Guidelines 2.0. Online, last accessed June 13, 2008, 2008. `http://library.gnome.org/devel/hig-book/stable/`.

[13] R. Hastie. Problems for judgement and decision making. *Annual Review of Psychology*, 52(1):653–683, 2001.

[14] E. Hutchins. How a Cockpit Remembers Its Speeds. *Cognitive Science*, 19(3):265–288, 1995.

[15] V. Kaptelinin. Activity Theory: Implications for Human–Computer Interaction. *Context and consciousness: Activity theory and human–computer interaction*, pages 102–116, 1996.

[16] V. Kaptelinin, B.A. Nardi, and C. Macaulay. Methods & tools: The activity checklist: a tool for representing the "space" of context. *Interactions*, 6(4):27–39, 1999.

[17] L.T. Kohn, J. Corrigan, and M.S. Donaldson. *To Err Is Human: Building a Safer Health System*. National Academy Press, 2000.

[18] S. Krug. *Don't Make Me Think!: A Common Sense Approach to Web Usability*. Que, 2000.

[19] L. Larsson. Guidelines. TDBB26 lecture notes, online, accessed June 13, 2008, 2006. `http://www.cs.umu.se/kurser/TDBB26/HT06/l02.pdf`.

[20] F. Lau. Toward a framework for action research in information systems studies. *Information Technology & People*, 12(2):148–175, 1999.

[21] Renaissance Learning. Accelerated Reader website. Online, accessed June 13, 2008. `http://www.renlearn.com/ar/`.

[22] A.N. Leontiev. *Problems of the Development of the Mind*. Moscow: Progress Publishers, 1981.

[23] R.P. Leow and K. Morgan-Short. TO THINK ALOUD OR NOT TO THINK ALOUD: The Issue of Reactivity in SLA Research Methodology. *Studies in Second Language Acquisition*, 26(01):35–57, 2004.

[24] H. Lindgren. *Decision Support in Dementia Care: Developing Systems for Interactive Reasoning*. PhD thesis, University of Umeå, 2007. UMINF 07.02.

[25] H. Lindgren. Collaborative Knowledge Building for Decision-Support System Development. To appear in proc. HCIS2008, 2008.

[26] H. Lindgren. Decision Support System Supporting Clinical Reasoning Process — an Evaluation Study in Dementia Care. To appear in proc. MIE2008, 2008.

[27] H. Lindgren and P. Eklund. Differential diagnosis of dementia in an argumentation framework. *Journal of Intelligent and Fuzzy Systems*, 17(4):387–394, 2006.

[28] R. Lipshitz, G. Klein, J. Orasanu, and E. Salas. Focus article: Taking stock of naturalistic decision making. *Journal of Behavioral Decision Making*, 14(5):331–352, 2001.

[29] J. Malm. T10: "Huvudkursen" – Hemsida. Online, last accessed June 13, 2008, 2008. `http://www3.umu.se/clin_sci/ophthal/schema/`.

[30] L. Mathiassen, A. Munk-Madsen, P.A. Nielsen, and J. Stage. *Objektorienterad analys och design*. Studentlitteratur Lund, 1998.

[31] B.J. McNeil, S.G. Pauker, H.C. Sox Jr, and A. Tversky. On the Elicitation of Preferences for Alternative Therapies. *Judgment and Decision Making: An Interdisciplinary Reader*, 2000.

[32] R. Moreno. Decreasing Cognitive Load for Novice Students: Effects of Explanatory versus Corrective Feedback in Discovery-Based Multimedia. *Instructional Science*, 32(1):99–113, 2004.

[33] G.E. Mueller. The Hegel Legend of "Thesis-Antithesis-Synthesis". *Journal of the History of Ideas*, 19(3):411–414, 1958.

[34] B.A. Nardi. Studying context: A comparison of activity theory, situated action models, and distributed cognition. *Context and Consciousness: Activity Theory and Human–Computer Interaction*, pages 69–102, 1996.

[35] J. Nielsen. *Usability Engineering*. Morgan Kaufmann, San Fransisco, 1994.

[36] J. Nielsen. Ten Usability Heuristics. Online, last accessed June 13, 2008, 2005. `http://www.useit.com/papers/heuristic/`.

[37] Pirkko Nykänen, Jytte Brender, Elske Ammenwerth, and Jan Talmon. *Guidelines for Best Evaluation Practices in Health Informatics*, 0.8 edition, 10 2007. Work in progress.

[38] C. Olsen. Automatic Assessment of Mammogram Adequacy. Licenciate thesis, Umeå University, 2005.

[39] F. Paas, A. Renkl, and J. Sweller. Cognitive Load Theory and Instructional Design: Recent Developments. *Educational Psychologist*, 38(1):1–4, 2003.

[40] V.L. Patel, J.F. Arocha, M. Diermeier, J. How, and C. Mottur-Pilson. Cognitive psychological studies of representation and use of clinical practice guidelines. *International Journal of Medical Informatics*, 63(3):147–167, 2001.

[41] V.L. Patel, D.R. Kaufman, and J.F. Arocha. Emerging paradigms of cognition in medical decision-making. *Journal of Biomedical Informatics*, 35(1):52–75, 2002.

[42] J. Preece, Y. Rogers, and H. Sharp. *Interaction design: beyond human–computer interaction*. Wiley, New York, 2002.

[43] H.A. Simon. Satisficing. *The New Palgrave: A Dictionary of Economics*, 4:243–245, 1987.

[44] S. Smith-Atakan. *Human–Computer interaction*. Thomson, 2006.

[45] S. Suebnukarn and P. Haddawy. COMET: A Collaborative Tutoring System for Medical Problem-Based Learning. *IEEE INTELLIGENT SYSTEMS*, pages 70–77, 2007.

[46] E. Toth-Pal. *Computer decision support systems for opportunistic health screening and for chronic heart failure management in primary health care*. PhD thesis, Karolinska Institutet, 2007.

[47] A. Tversky and D. Kahneman. Judgment under Uncertainty: Heuristics and Biases. *Science*, 1974.

[48] S.W. van den Braak, H. van Oostendorp, H. Prakken, and G.A.W. Vreeswijk. A critical review of argument visualization tools: Do users become better reasoners. *ECAI-06 CMNA Workshop*, 2006.

[49] J.J.G. van Merriënboer, P.A. Kirschner, and L. Kester. Taking the Load Off a Learner's Mind: Instructional Design for Complex Learning. *Educational Psychologist*, 38(1):5–13, 2003.

[50] L.S. Vygotsky. Mind in society: The development of higher psychological processes. *Cole, V*, 1978.

[51] C. Wilhelmsson. Lecture. Lecture notes from Medical training at Umeå University, 2008.

[52] A. Wimo, L. Johansson, and L. Jönsson. Demenssjukdomarnas samhällskostnader och antalet dementa i Sverige 2005. Socialstyrelsen article number 2007-123-32, 2007.

A Fabricated patient case

The following text, in Swedish, was used as the fabricated patient case in the evaluation study. The text itself was written by Helena Lindgren. The text is split in two parts, where the first part was shown immediately and the other was shown first after having used the system.

> Per (75 år) har successivt blivit sämre vad gäller minne och orienteringsförmåga, enligt makan. Han går vilse i sin egen skog som han arbetat i de senaste 50 åren. Han tappar bort saker och sätter igång projekt som blir oorganiserade och ibland farliga när gårdens maskiner ska användas. Per är ofta uppe på nätterna och har saker att göra, och ofta upplever han att det är bråttom av någon anledning som han inte kan redogöra för. Per märker att något inte fungerar och blir stressad och aggressiv när han upplever att anhöriga gömmer saker för honom. Per kopplar själv en del till att minnet inte är som förut. Makan är trött och orkar inte hålla samma takt, och är rädd att det ska hända en olycka. Pers välkända historier om gamla tider blir allt fattigare och ofta behöver Per tänka efter för att minnas detaljer. Han ställer saker på fel ställen som exempelvis pappersrullen i kylskåpet, vad gäller ADL får han fundera länge för att få på sej kläder i rätt ordning, men de är ofta ut och in. Saker som förut passade ihop gör det inte längre, stickproppar passar inte i vägguttagen, mm. Per är dock verbal, oftast glad och lyckas enligt makan ofta prata bort saker som blir fel. De aggressiva inslagen upplevs dock främmande och makan tycker inte Per är sej lik vissa stunder. Per tycker sej dock inte se saker som inte finns, vilket stämmer med makans uppfattning. Symptomen uppges inte variera över dygnet.
>
> Förutom en galloperation 10 år tidigare noteras inga tidigare sjukdomar, ej heller några toxiska expositioner eller alkoholmissbruk. Labbprover UA. Likvorprov ej tagna.
>
> MMT 21/30
>
> Klocktest 0
>
> Remiss Neuropsyk

The second part follows:

> *NEUROPSYK* klinisk intervju och USK med delar ur WAIS-III
>
> Orientering: Klart under normalzon
>
> Verbal förmåga: God
>
> Logiskt tänkande: Ua
>
> Visuoperceptuell och spatial förmåga: klart under normalzon

Uppmärksamhet och psykomotorik: nedsatt, har särskilt svårt att klara uppgifter som kräver samtidig förmåga till uppmärksamhet, flexibilitet och snabbhet.

Exekutiv förmåga: klart under normalzon

Minne och inlärning: Episodiskt minne nedsatt för både verbalt och ickeverbalt material.

B Questionnaire

The questionnaire for investigating the attitudes toward the system was written in Swedish. For convenience, these have been translated to English here. The original text in Swedish is in parentheses. For each statement, the test subject was asked to mark the appropriate answer in a box. The possible answers to the questions were:

- Do not agree at all ("instämmer inte alls")
- Agree to some extent ("instämmer till viss del")
- Agree to large extent ("instämmer till stor del")
- Agree completely ("instämmer fullständigt")

The statements were as follows:

- The system affected which investigation steps I would perform with the patient ("Systemet påverkade vilka undersökningar jag skulle genomföra med patienten")
- The system affected which treatment I would recommend for the patient ("Systemet påverkade den behandling jag skulle rekommendera för patienten")
- The system may act as important support for my future activity ("Systemet kan agera ett betydelsefull stöd för min framtida verksamhet")
- The system was easy to use ("Systemet var enkelt att använda")
- I would recommend the system to my colleagues ("Jag skulle rekommendera systemet till mina kollegor")
- I feel positive toward the system ("Jag är positivt inställd till systemet")
- I want to use the system, or a similar one, in the future ("Jag vill använda systemet, eller liknande, i framtiden")

C Results from questionnaire

The results from the attitude questionnaire in the main evaluation study are presented below.

C.1 The system affected which investigation steps I would perform with the patient

- Do not agree at all: 1
- Agree to some extent: 5
- Agree to large extent: 1
- Agree completely: 1
- Blank: 1

C.2 The system affected which treatment I would recommend for the patient

- Do not agree at all: 1
- Agree to some extent: 4
- Agree to large extent: 2
- Agree completely: 0
- Blank: 2

C.3 The system may act as important support for my future activity

- Do not agree at all: 0
- Agree to some extent: 2
- Agree to large extent: 4
- Agree completely: 3
- Blank: 0

C.4 The system was easy to use

- Do not agree at all: 0
- Agree to some extent: 2
- Agree to large extent: 6
- Agree completely: 1
- Blank: 0

C.5 I would recommend the system to my colleagues

- Do not agree at all: 0
- Agree to some extent: 2
- Agree to large extent: 6
- Agree completely: 1
- Blank: 0

C.6 I feel positive toward the system

- Do not agree at all: 0
- Agree to some extent: 0
- Agree to large extent: 6
- Agree completely: 3
- Blank: 0

C.7 I want to use the system, or a similar one, in the future

- Do not agree at all: 0
- Agree to some extent: 1
- Agree to large extent: 6
- Agree completely: 2
- Blank: 0