# How Humans solve Scheduling Problems:

## Analysis of Human Behavior in the Plan-A-Day task

**Diploma Thesis of Stefani Nellen**

Ruprecht Karls Universität, Heidelberg
Department of Psychology

Date of Submission: April 2002

Advisor and first Reviewer: Prof. Joachim Funke
Second Reviewer: Prof. Marcus Spies

Stefani Nellen
Hans-Thoma-Str. 72
69121 Heidelberg
Tel.: 06221/ 373510
Email: Stefani.Nellen@urz.uni-heidelberg.d

**Abstract**

This thesis explores features of the Plan-A-Day (PAD) task by Funke & Krüger (1993) and presents an analysis of specific aspects of scheduling behavior.

The PAD task permits learning at a declarative as well as on a procedural level. Declarative learning in the PAD domain can be conceptualized as the accumulation of experience about the feasibility of partial schedules. An "explorative pattern" is defined which characterizes a scheduling process that implements the strategy of accumulating experience. Procedural learning is hypothesized to take place at the level of the mental arithmetic that is necessary to check schedules in advance (forward checking).

It is shown how declarative and procedural learning must work together to enhance scheduling in the PAD task.

These assumptions are further investigated in two studies. In the first study, the "explorative pattern" defined in this thesis is confirmed by the analysis of empirical data. In this study, it also showed that participants who explored little in their first PAD session performed worse in the second session, while the performance of participants who explored much during the first trial improved slightly. Furthermore, in this study a considerable increase of forward checking between the two PAD sessions was found, confirming the assumption about forward checking being the basic method in working with PAD.

In the second study, participants evaluated partial schedules. It showed that the feasibility of other appointments is an important reason for evaluating partial schedules. However, according to this study, the actual choice of an appointment is less related to forward checking but to criteria of the appointments. This is interpreted as a qualification of the role of forward checking as the basic skill underlying performance in PAD by implying that there is a preliminary selection process that precedes it.

## Acknowledgements

Although I figure as the sole author of this thesis, and consequently have to face any criticism concerning its content on my own, I am happy to use this opportunity to give thanks to the people that helped and supported me throughout its composition.

Firstly, I want to thank Prof. Joachim Funke for being a stimulating, challenging, and yet always patient and supporting advisor. I am especially grateful to him for introducing me to the field of Cognitive Modeling and the ACT-R architecture, which, although not explicitly featured in this work, have proved to be most fortunate influences in both my scientific and my private life. Similarly, I want to thank Prof. Marcus Spies, my second advisor. Although our meetings were, due to his current work in Munich, restricted to a minimum, I found his feedback from the perspective of applied psychology and programming most inspiring.

Two people need to be mentioned here whose scientific competence and support provided me with insights I wouldn't have been capable of on my own. They are Jon Fincham from Carnegie Mellon University (Pittsburgh), and Jan Reimann, from the Department of Mathematics (University of Heidelberg).

Jon was my mentor during the ACT-R Summer School in the year 2000. At this time I was trying to develop a Cognitive Model for the Plan-A-Day task, which, for reasons that are outlined in chapter 3 of this thesis, proved to be terribly hard. Jon saw an early version of my model through a very dark night, and, during the "two weeks of madness" in the summer of 2000, also became a good friend. Without him, I probably wouldn't do science anymore now, so it's all his fault!

Jan was my advisor for a strange presentation on decidability I did for an even stranger course in formal languages. Due to his rare combination of mathematical and verbal skills, the discussions I had with him in the course of preparing my presentation (and afterwards) were not only a pleasure, but they also provided me with insights that easily surpassed anything I was able to get out of cognitive-science literature. I hope I have learned enough from him about clear explanations of difficult things to make this thesis readable to psychologists and non-psychologists alike.

I would hardly have survived the writing of this thesis in full possession of my sanity without the support and presence of my friends Sonja, Marlis, Jan-Ulrich and Timm Lochmann.

Sonja Schildheuer, known to the world of aviation as Sonja O'Leary, is the author of "How low can you go" -a brilliant dissertation in international business administration, which, incidentally, was composed at the same time as the mediocre thingie you are holding in your hands. Young maiden, I probably would have made it without you, but certainly without the style appropriate for Air Force wives. Thank you for quenching the persistent temptation to send that letter, starting "I'd rather

Marlis Schlüter has been an expert at the gentle art of distraction and (of course) time-planning (she should have written a thesis on scheduling, not me!), touchingly luring me to spend the occasional evening in town, or at her deliciously prepared dinner table. Escape to Life!
Jan-Ulrich Schmidt, although he has kissed psychology goodbye in favor of a life in art long ago, has heroically refrained from mocking me during the composition of this work and has never ceased to believe in me and my subsequent travels to the U S. Pittsburgh!
Timm Lochmann shared many mad caffeinated conference experiences with me (long live Assen), and always remained the (mysteriously) calm eye in the center of the scientific storm. Good Luck with Mr. Boltzmann!

I also want to thank my mother, Moni, and my aunt, Margit, for always being supportive, loving, and apparently unfaltering in their belief in my capabilities. That meant very much to me.

The final spot of any acknowledgements is traditionally reserved for the author' s partner, so you are certainly not surprised, Niels, to find your name here...and what can I tell you, apart from the things that would require pages and pages, but are, unfortunately, out of place in an academic thesis! Let me start by thanking you for offering me a refuge in Groningen when I needed to get away from Heidelberg, and for unfailingly pointing me to the right argument/ paper/ article when I felt that I had run out of ideas. Thank you for supporting me without putting me under pressure, for inspiring me and for distracting me, for being there for me now and for dreaming with me about the future. I love you.

Stefani Nellen

Heidelberg, 05.04. 2002

## Contents

# 1. Introduction and Objective of this thesis

The concept of planning is neither unambiguous, nor narrow. Consider a simple field study (carried out by author, imaginary). If ten people were picked randomly and asked about their first associations, given the term "planning", their answers would be likely to range from dreamy confessions in the style of "My husband and me are planning to move to Florida" to brusque statements like "I was just planning to eat this sandwich here when you came and interrupted me with your question!"

However, some people would probably answer simply by giving details of the activities they have planned to do during the present day, as this hypothetical student does:
"Well, I planned to go shopping, but only for an hour or so, because afterwards I'll meet a friend at the café. Perhaps we'll do something else afterwards. But I have to be home early tonight, at 8 in the latest case, because tonight I have to prepare a presentation which I have to give tomorrow, which means I will go to bed rather late and sleep rather little. The presentation will be tomorrow morning, and if it goes well, I'll have a little celebration afterwards".
This last type of "planning", with the flavor of "love in idleness" that is so typical of the life of students, closely resembles the type of planning that is operationalized in the task that is used in this thesis: the Plan-A-Day task (Funke & Krüger, 1993).
The Plan-A-Day task is, as its name already suggests, about the scheduling of several activities during a day. These activities come with several constraints: They can be met only at specific times, or between specific points of time. This is also true for the schedule our imaginary student has described. The meeting with the friend is scheduled for a specific time, and the student wants to be home "at 8, in the latest case".
There is another element of scheduling, which is not mentioned explicitly in the statement above, but nevertheless is of considerable importance. This is the distance between the locations the appointments take place at. E.g., our student may have arrived at an estimate for the time that is available to do shopping by calculating the distance from the shop to the café, where the friend will be waiting (and the distance from his/her current position to the shopping destination as well). The appointments in the Plan-A-Day task also have distances between them. They are also assigned specific priorities, i.e. they are not all equally (un-) important. The fact that different appointments do have different priorities can also be

restrict the duration of shopping to an hour. However, the presentation scheduled for the following morning appears to be of even higher importance, as its success will be explicitly celebrated. Furthermore, the student will probably discard after-coffee fun in town with the friend in favor of properly preparing the presentation: another indicator of the top-level priority of this "appointment".

What about the agonizing situation when two apparently equally important appointments take place at approximately the same time, and there is no way to meet them both? Well, that situation can also occur in the Plan-A-Day task. The task (or rather: the Interface, as Plan-A-Day is of course implemented to run on a computer) does nothing to help participants come to term with that that dilemma, as it does not yet contain a first aid therapeutic facility. All it can do is to record participants' eventual decisions in a log-file, which is interesting for the researcher, but offers no real consolation to the participant.

Let me briefly interrupt my description of the Plan-A-Day task and give my reasons for stressing its realistic features.

I hold (or rather: share) the view that any psychological (or, indeed, scientific) work concerning itself with planning and scheduling that uses a specific task to assess human behavior in these areas has to be very clear about the nature of this task. And this is also true for this thesis. This is because the terms "planning" and "scheduling" are often used to refer to tasks, phenomena and findings that are considerably different from each other, which can lead to false generalizations and erroneous "contradictions". A (very) brief sketch of the concepts of planning in the fields of Artificial Intelligence and Psychology may help to clarify that point.

Hertzberg (1989, 1995) describes the development of the concept of planning in Artificial Intelligence (AI). He mentions several problematic characteristics of real life planning and scheduling domains. Among these characteristics are (e.g.) the dynamic nature of many tasks, the necessity to find representations for the passage of time to be included in planning and scheduling algorithms, or the fact that planners often have do deal with incomplete information. These phenomena have led to a considerable diversification of the concept of planning in AI, because one of the main objectives of that field lies in developing an efficient planner for a specific domain, and as the demands of that domain change, so do the characteristics of a planner. Hertzberg (1995) concludes: "There is no such thing as planning *in general*" (p. 91) and proposes the thesis that "models of planning and to represent the

characteristics they imply" (p.92). He goes on to demand that "each model must explicitly state its definition of a plan and of a problem" and clearly define its "Operatorkalkül" (a German term that is a bit hard to translate, meaning "the mechanism which is used to evaluate the consequences of specific components of a plan, or the complete plan", p.92).

As planning and scheduling are research areas that have proved to be similarly challenging for AI and psychology (see e.g. Akyürek, 1992; Funke & Fritz, 1995a; Hayes-Roth & Hayes Roth, 1979; and Rattermann, 2001), it comes as little surprise that the diversification that emerged in the field of AI-planning is also prevailing in the field of psychology. As Sanderson (1989) states in a survey on human scheduling in the domain of job scheduling and dispatching, the variety of tasks that are used assess human scheduling capabilities is vast enough to make it impossible to classify studies of human scheduling according to the task they use. Instead, Sanderson (1989) proposes more abstract criteria, e.g. the level of scheduling expertise in the human sample that was assessed, or the amount of information available to the participants.

Although this last point was made with regard to scheduling in the industrial/ human factors domain, which, strictly speaking, is not equivalent to "pure" psychology, it nevertheless applies to the situation in psychology as well. The tasks used in psychology can be (outwardly) simple and context-free tasks like the Tower of Hanoi (Klix & Rautenstrauch-Goede, 1967) or the Tower of London (Shallice, 1982). However, the family of scheduling tasks also encompasses close relatives or even siblings of Plan-A-Day (e.g. the "A day's errands" task used by Hayes-Roth & Hayes-Roth (1979)). Finally, there are still more complex scenarios mostly used in Dynamic System research, where planning is only one among many relevant variables (see Funke, 2001 and Wallach, 1998, for overviews).
To add to the complexity of the planning/scheduling picture, the same task is often used in different fields of psychology. E.g., the Tower of London and the Tower of Hanoi were both used to assess planning impairments of patients suffering from prefrontal damage (for a comparison, see Huchler, 1999). On the other hand, both "Towers" were (and still are) of course widely used to assess basic cognitive processes in healthy adults (drawing an amusing analogy to molecular genetics, Herbert Simon (1996) referred to the Tower of Hanoi as "the *e. -coli* of cognitive psychology" (chess being the *drosophila*) (p.226)). The emergence of planning competence is also of interest to developmental psychologists, as can be seen in a

recent work by Rattermann et al. (2001), which uses a task that resembles Plan-A-Day to investigate the emergence of partial-order planning in children.
Finally, the extensive use of scheduling tasks in the field of personnel selection need not be specially emphasized here, as it is well known (but see Funke & Fritz, 1995b, for a brief overview).

In spite of their necessarily superficial character, the preceding paragraphs should suffice to show that the collected research on planning and scheduling does indeed resemble a colorful jungle full of diverse interesting specimens of scientific flowers, where only one thing seems to be impossible: To answer the question that so boldly constitutes the title of this thesis, "How humans solve scheduling problems". "Humans" could be just adult, healthy humans, but the term also should include children and neurological patients. "Scheduling problems" can be based on realistic scenarios, like scheduling in Plan-A-Day, but they could also be highly specific problems like the (simulated) control of a nuclear plant (cf. Wallach, 1998), or, very purely and abstractedly, amusing little puzzles.
So, how is the diversity-problem[1] dealt with in this thesis, then?
One possible approach to introduce order into the "jungle of planning" ("Planungsgestrüpp", Funke & Fritz, 1995a, p.37) could be the design of a taxonomy of planning tasks and different kinds of planning. Indeed, Funke & Fritz (1995a) offer some tentative ideas about the dimensions that could be part of such a taxonomy. However, this thesis uses a different approach, the essence of which can be found in its subtitle: "Analysis of human behavior in the Plan-A-Day task". In other words, I chose to concentrate my investigations of scheduling on a specific task. This approach is in accordance with Newell's (1973) suggestion to learn as much as possible from a single task. Such a choice is legitimate if

- the definition of the problem/ phenomenon *within the task* is clear and the characteristics of the task (e.g. its difficulty) are made explicit[2] and
- the task is realistic enough to cover at least a segment of the phenomenon in question that could occur in real life.

And here we are again at the point from which we started this excursion, namely, the importance of the realistic features of the Plan-A-Day task: They are important, because, within the terminological boundaries of this thesis "scheduling" equals "Plan-A-Day-

[1] Or rather, less negatively, "phenomenon" or even "challenge"

[2] This is an explicit response to Hertzberg's (1995) demands for clarity of definition and specification quoted

behavior", but it would be nice if "Plan-A-Day-behavior" at least partially equaled real behavior.

So "How humans solve scheduling problems" should actually be read as "how healthy, adult humans work with a scheduling task called Plan-A-Day", and contain a generous amount of task analysis.

However, I chose to focus my ambition to describe human scheduling in yet another way. In this thesis I want to look at three specific aspects of the scheduling process. Firstly, I am going to investigate how often and in what ways humans modify a schedule until they are "satisfied", and arrive at their solution. Do they make small, "local" changes to a schedule, gradually optimizing it? Or do they tend to abandon schedules quickly, and take them up again frequently?

Furthermore, I want to explore the extent of "looking ahead" that is employed in a task like Plan-A-Day. How many steps in advance do people detect that the schedule they are currently working on will lead to a dead end, i.e. that it will be impossible to include all the appointments that have to be met?

A third objective of this thesis is to explore how people evaluate single appointments. If a set of appointments is given, which appointments are evaluated to be good choices to start a schedule with, and which are not? And which criteria are critical for this evaluation? How do these evaluations correspond to the actual choices made by other people?

In focusing on these three aspects of scheduling, I follow suggestions made by Funke & Krüger (1993, 1995) in the course of their description of the Plan-A-Day task. The character and the extent of schedule-modification and "look ahead" are mentioned explicitly as plausible additional measures of the planning process that should be computed along with those already provided by the Plan-A-Day system (Funke & Krüger, 1993, p. 108)[3]. The question of how humans evaluate partial schedules is mentioned as an additional option to diagnose planning capabilities (Funke & Krüger, 1995, p.118).

The fact that the (re-)construction of schedules, the extent of look-ahead and the evaluation of single appointments or partial schedules were deemed worthwhile research-pursuits by the authors of the task I worked with was an essential and important inspiration that guided my analysis of the empirical data I collected in the course of writing this thesis. However, it was

not the only thing about these three phenomena that attracted my curiosity. I was particularly fascinated by the ambivalent nature of modifying a schedule and looking ahead in time – "ambivalent" because they both are a necessary part of efficient scheduling, but can change into a real obstacle when they are "overdone". Consider the following example.

If I have, say, six appointments today, I must come up with a schedule to meet them, because for this number of appointments, purely spontaneous behavior ("oh, let's go to the Conference first, it sounds like fun") can result in considerable loss of time and, subsequently, stress. So I start building a schedule, using the information about the allotted time for the appointments and the distances between them (perhaps I have a map, or someone told me, or I know my way around in the city). With six appointments, it is unlikely that I'll come up with the right schedule immediately, especially since this scheduling involves a lot of mental arithmetic, computing time estimates, etc., and I'm not really superior at that. Now suppose I have to write down that schedule, either for my own use (to take it with me, so I wont forget it on the way), or for somebody else, maybe because I work as an assistant for somebody who is so important that they can't be bothered with something as trivial as scheduling their daily activities. Anyway, I have two options. I can either start sketching a tentative schedule, with the danger of having to correct it later, crossing out appointments, making things a bit messy, or carefully, carefully think about the appointments until I come up with a schedule that certainly works, and write it down neatly and clearly and in one go. What should I do?

This question is really about the most appropriate extent of schedule modification, the recommended number of steps to look ahead during scheduling, and the consequences of both.

Let's address the look-ahead question first. Assume I choose to adopt the "think carefully" -strategy mentioned above. In this case, I can be lucky and arrive at a complete schedule by accident. However, it is more likely that I have to think four or five steps into the future to be really certain that the schedule will in fact work out. This not only sounds very straining, it is also likely to lead to calculation errors, because I have to keep track of so many things at once: The current time, the alternatives to the current appointment, the differences and sums of the various times. No, looking ahead too far is not recommended, as it is hard work that is not even guaranteed to succeed. However, the other extreme, a very limited amount of looking-ahead (i.e. simply adding an appointment to the schedule that can be done at that time) also has its pitfalls. While I can, again, strike it lucky and arrive at the complete schedule for my appointments in time, I am now in danger arriving at a dead end

unnecessarily often, because I haven't seen it coming in time. Moreover, I'm now more likely to have to repair large parts of my schedule, because I may have overlooked an appointment that started very early, and could only be met until a relatively early time, and I have to insert that appointment into the beginning of the schedule. It is easy to see that scheduling without looking ahead is inefficient. However, too much looking ahead is a strain. Looking ahead is useful only in moderation.

The question of modifications to a schedule can be answered in a similar vein: Like with looking-ahead, moderation is the key here. Basically, I shouldn't modify a schedule without a good reason, e.g. without knowing for sure that I can't meet some appointment. However, who doesn't know the sudden urge to change a schedule and see how things work out when a different appointment is placed first. These modifications can be useful, as they help to avoid the danger of being stuck. Remember Francis Picabia, who said: "Why is the skull round? To enable thoughts to change direction!" But too many modifications to a schedule can not only be a sign of severe problems of the scheduler, they are also a poor strategy. If I switch between schedules starting with different appointments too often and too quickly, I cannot collect knowledge about the appointments; I can't accumulate experiences about sequences of appointments that are possible. However, these bits of knowledge would make things considerably easier for me (which is probably another reason for the above-described urge to "change directions" occasionally). But if I modify *one single schedule* too often, I will be stuck and persevere on a road that leads nowhere – perhaps only a tiny modification away from the solution.

This necessity to maintain a delicate balance both in modifying a schedule and in looking ahead in time interested me, because it so close to a notion of common sense, of flexible human scheduling, as opposed to the mind-numbing search tree routines of algorithms[4]. And so I wanted to take a closer look at *just how* humans modify their schedules.

Let me now give a brief overview of the remainder of my thesis.

In the following chapter I will describe the Plan-A-Day task in more detail, the task environment and Interface as well as the ways in which scheduling abilities are assessed by

[4] A rather unwarranted generalization made in the flow of polemics. In fact, there are many algorithms that are sophisticated and not mind - numbing. A funny example is the "Dynamic Backtracking" Algorithm in Ginsberg

the automatic evaluation processes implemented in the system. I will also introduce the terminology used in the course of my own analyses throughout this thesis.

Task analysis will continue in chapter 3. I will discuss classical planning and Constraint Satisfaction Search, as well as three ”prominent” models of planning, the design of each reflects a “growing concern about cognitive plausibility” (Akyürek, 1992, p.82). I’ll discuss the memory-driven approach of Adaptive Planning (Altermann, 1988), inspired by Schank & Abelson ’s (1977) script-theory, and Opportunistic Planning, a hybrid idea of Psychology and AI, as well as an early example of Cognitive Modeling (Hayes-Roth & Hayes-Roth 1979). I will also review Anderson’ s (1987) procedural view of skill acquisition (based on his ACT* theory, 1983). I will use all these concepts in order to carry the “superficial” task analysis of chapter 2 a little further by critically evaluating their appropriateness in the Plan-A-Day context. From this assessment, I will deduce a proposal about the connection between declarative and procedural learning in the domain of PAD, or, to put it more informally, between the accumulation of experience and the improvement of performance.

After these task-analytical and theoretical musings, I will address the question of different patterns of modifications in human scheduling behavior in chapter 4. The analysis presented in this chapter partly derives from the ideas expressed in chapter 3. Apart from the obviously interesting question “how many modifications do participants “need” before they arrive at their solution”, it can also be analyzed why some participants take longer (in terms of modifications) than others. Is it because they can’t stop working with a single schedule? Is it because they take up the same schedules over and over again, being stuck? Is it because they try out too many different things, without actually succeeding? And, finally, is there something like a “scheduling style”, i.e. are participants who take many modifications in a first Plan-A-Day task likely to take long in the second task as well? Or is it rather the scheduling-patterns that are consistent across the different tasks? Put a bit more casually: are the modifications systematic or scattered?

The phenomenon of “looking-ahead” will be addressed in chapter 5.
In this chapter I will describe an experiment I conducted to test how people evaluate single appointments at the start of a schedule. This experiment uses the same Plan-A-Day tasks (i.e. the same appointments) that were used in the study described in chapter 4. This makes it easy

given in the experiment, which was one of its objectives. However (more interesting), people in the experiment were also asked to give reasons for their evaluations. It will therefore be possible to test if these evaluations are really rooted in participants' assessment of the future situation (looking-ahead), or if other, more simple criteria are enough.

In the experiment, people were also asked to pick the next appointment, given a schedule starting with a specific appointment, and to give their reasons for this as well. This further differentiates participants' reasoning in evaluating the next appointment, because the evaluation of (another's) already existing schedule is not quite the same as building a schedule from scratch. Looking-ahead may play a role in the evaluation of other's efforts, while simple priority rules (cf. Sanderson, 1989) may be enough to choose the next appointment.

Finally the findings are summarized in chapter 6, focusing on the question that perhaps would make the best title for this thesis after all: Is human scheduling any good?

# 2 The Plan-A-Day Task

This chapter contains a detailed description of the task that is used to assess scheduling throughout this thesis: The Plan-A-Day task, or PAD for short (Funke & Krüger, 1995). The external features of the task are described, e.g. the characteristics of the Interface, the amount of information that is available to the subjects, the setup of the situation the subject is placed in, and the wording and contents of the instructions. This description should provide the reader who is unfamiliar with the Plan-A-Day task with a clear concept of what it is like to deal with the task as a participant.

In addition, the format of the log-data collected by the PAD-System will be discussed shortly, in order to introduce properly the terminology that will be used in the subsequent chapters to describe, explain and predict human scheduling behavior. Finally, this section also includes a list of abbreviations for the appointments featured in two specific PAD tasks, because the subsequent chapters will extensively refer to these two specific tasks.

## 2.1 Development of PAD and special features of the task

PAD was developed by Funke and Krüger (1993), originally with the purpose to devise a diagnostic instrument for the assessment of planning and scheduling capabilities of executive personnel. However, a special version of PAD, the "PAD-Reha", was designed explicitly as a means to make a diagnosis for patients with neuropsychological deficits (see Huchler, 1999, as well as Kohler, Poser & Schönle, 1995, for an evaluation).

Both the development of the "Standard" -PAD and the PAD- Reha result from the authors wish to extend and improve diagnostic instruments that already exist in the area of scheduling. PAD is very similar to an earlier "Disposition-task" by Jeserich (1981). In this task, several places have to be visited in the course of an afternoon, e. g. the grocery-store (to buy food), the doctor (to have a routine health check), and the hairdresser (for obvious reasons). The participant in this task has to order these appointments in a way that enables them to meet them all. A bicycle can be used once, to reduce the distance between two appointments to a third, however, as this bicycle is broken and has to be repaired first, the use of this device also requires additional time and, as such, must be considered carefully.

As Funke and Krüger (1993) argue, PAD, while maintaining the basic framework of this Disposition-task, improves it in several crucial ways: Firstly the appointments in PAD are more similar to the appointments one is likely to encounter during a working Day. Instead of buying milk at the grocery store, the participants have to (e.g.) attend a Conference, dictate a letter to the Secretary and meet their boss at the Central Office. Secondly, the different

appointments have different priorities, also a familiar feature of "real-life" obligations. The different priorities of the appointments are also used to qualify the evaluation of the schedules participants produce in the PAD task. Sometimes, it is not possible to meet all scheduled appointments for one day, which means that the participants have to select a subset of appointments they want to meet. This is another realistic feature of the PAD task, as the difficult and crucial aspects of scheduling often lie in deciding between two or more conflicting appointments. A third modification from the Disposition task is the "exchange" of the bicycle in favor of a car. Like the bicycle in Jeserichs (1981) original task, the car can be used once each day, and reduces the distance between two appointments to a third.

A fourth group of modifications serves to enhance PAD's diagnostic quality: The participants are required to schedule appointments for two (instead of (only) one) days, in order to improve the assessment of their scheduling abilities by means of repeated measurement. Furthermore, the difficulty of the individual "days" (i.e. the number of appointments and the conflicts between them) can easily be changed according to specific research interests. Even the words describing the appointments themselves can be changed in that manner. Finally, PAD not only provides a measure of the participants' performance, i.e., the results of their scheduling. It also provides a means to analyze the scheduling process itself, because it generates a log-file for each participant. This log-file holds every keystroke the participants make. This enables the interested researcher, such as the author, to address specific questions about the scheduling process, in addition to the quality of the results (i.e. the final schedules participants create). (An account of the ways human scheduling can be evaluated by the PAD system will be given in section 2.4. However, as most analytic procedures that were used in this thesis were not a part of the default options already implemented in PAD, but were instead programmed by myself[5] for the specific purposes of this work, this account will be not as thorough as the subject matter would warrant. However, Funke and Krüger (1993) give a very clear and detailed description of the evaluation options provided by PAD.)

[5] With one notable exception that has to be mentioned here: The data described in chapter 3, section 3.1 were obtained by using a program ("Einles") written by my fellow student Jan Zwickel, for whose help and assistance

## 2.2 Options for the configuration of PAD

PAD was written in Turbo Pascal 6.0 and runs under MS DOS or Windows[6]. Along with PAD come 16 pre-installed numbered sets of appointments. These sets of appointments differ with respect to their size. Additionally, there is one set of appointments ("0") that is used as an "exercise" and is presented to the participants prior to their regular work with PAD, in order to familiarize them with the system, the use of the correct keys, the demands of the task, etc.

It is possible to configure PAD to suit specific research interests and/ or the characteristics of the population whose scheduling behavior is to be assessed (e.g. healthy participants vs. neurological patients).

PAD's configuration options include the number of the two sets of appointments one wishes to present to the participants, as well as the "difficulty" of the task. This last parameter can be varied from level 1 (easy) to level 4 (very difficult). The difficulty as conceptualized as the presence/ absence of helpful information that is available to the participants while they solve the task. This information is designed to reduce the load on the participants' memory. E.g. the times at which the appointments can be met may be shown explicitly on the screen during the PAD-session. On difficulty level 0, all helpful information is available, on level 4 none. This point will be expanded a little more in the next section, where the actual PAD Interface will be described (and shown).

Other configuration options are the amount of time participants are given to schedule the appointments, and the turning on/off of a sound that warns them if their allotted time is about to run out. It is also possible to specify how many minutes prior to that moment the warning should occur. A last parameter is the running time; it can be specified if the passage of time is shown to the participant at the top of the screen or not.

After these technical details, we can launch into the PAD task, as a participant experiences it: the description of the PAD Interface.

---

[6] Or, as was the case in the experiment discussed in chapter 5, under Mac OS 9, using Virtual PC 4.0.

## 2.3 Scenario: A day in the PAD-World[7]

At the beginning of a PAD session, the participants sit down in front of the Computer and enter their name, age and gender (of course they do not have to enter their real names). After that, they are presented with the instruction, which can be summarized thus: The participants are asked to imagine themselves as employee of a company, who has to meet a number of appointments during a fictitious day. They are encouraged to meet as many appointments as possible. The appointments all take place within the area of the company, which consists of several buildings that are scattered over a wide area.

Participants are informed that each appointment can only be met at a specific time, or in a specific "time frame". They are also prompted to the fact that the scheduling of these appointments must take into account the distances between the respective locations. The option to take the car for one distance is mentioned.

After that, there follows an explanation of the possibilities to move between the locations in the PAD Interface by holding down the key that bears the first letter of the destination. Participants are told that they can always view the set of appointments they have to schedule, as well as general help, by holding down function keys. The option to delete moves and modify schedules is mentioned as well.

Now, the participants are presented with the exercise-trial that precedes the actual testing. This exercise consists of three appointments that have to be scheduled. Although the schedule itself is not hard to find, it involves the correct use of the drive-by-car option, which serves to prompt participants again at the importance to use that strategic device correctly. Only after having found the correct schedule are the participants allowed to enter the "regular" part of the PAD-Test. As already mentioned, it consists of two "days" for which appointments have to be scheduled.

This may be the appropriate moment to introduce the two sets of appointments that were used to obtain human data, in the study described in chapter 4 as well as in the experiment described in chapter 5. In the PAD system, they bear the numbers 4 and 5, so they will from now on be referred to as PAD 4 and PAD 5. The instructions for PAD 4and PAD 5 are shown

---

[7] I first read he term "PAD-World" in the Diploma thesis of my fellow student Wolfram Schenck (2001), which presents a connectionist model of planning in the domain of PAD. There, the term "PAD world" is used to describe the PAD as a kind of Microcosm, with its own definitions, terminology and cause-effect-relations. Formulations like "in the context of the PAD task", "within the Domain of PAD", are synonymous, but not half

in figure 2.1. As the analysis of these two sets of appointments is a crucial part of this thesis, and accordingly requires considerable space and elaboration, the differences between PAD 4 and PAD 5 (or, indeed, their characteristics) will not be commented upon here, but, instead, be analyzed more thoroughly in chapter 3.

PAD 4

- You have to be at the Storehouse between 10.00 a.m and 0.15 p.m. It will take you 10 minutes. *It's important.*
- Between 11.00 a.m. and 4.00 p.m you have to visit the Secretary. It will take you 10 minutes.
- You have to be at the Conference at 1.00 p.m, in the latest case. The Conference will last until 2.00 p.m. *It's very important.*
- You have to be at the Administration building at 2.30 p.m. It will last 90 minutes. *It's very important.*
- Between 10.00 and 4.00 p.m., you have to be at the Printing Office. This will take you 90 minutes. *It's very important.*

PAD 5

- Between 1.30 p.m and 2.30 p.m., you can meet a customer at the cafeteria. The talk will last 30 minutes. *It's important.*
- Between 11.00 a.m. and 14.00 p.m you have to show up at the Office and deal with the files there. You will need 60 minutes for this. *It's very important*.
- You have to be at the Conference at 11.30 a.m, in the latest case. The Conference will last until 0.15 p.m. *It's important.*
- Between 10.00 a.m. and 4.15 p.m. you have to meet your boss at the Central Office. He wants to see you for 10 minutes. *It's very important.*
- Between 10.00 a.m and 4.00 p.m., you have to be at the Administration. The work there will take 55 minutes. *It's important.*
- Between 10.00 a.m. and 3.00 p.m. you are to come to the Printing Office and copy a book. This will take 10 minutes

Figure 2.1. Descriptions of the appointments as they appear to the participants. Of course, participants are presented with one set of appointments at a time.

After having read the instructions, the participants may enter the actual PAD environment.

in minutes) between them. The subjects co-ordinate the subtasks by "moving to" the respective locations (as already mentioned, they do this by typing the first letter of the destination).

Each move results in a change of PAD system time, reflecting real-time relations and discrepancies between the subtasks. The subjects are allowed to delete and modify their moves, and declare their schedule finished, at any time. After that, they switch to the next "day". If a subject hasn't decided on a final schedule, this switch occurs automatically after fifteen minutes (there are two announcements that "time is running out" before that).

Let's take a closer look at the PAD-interface, which is shown below.

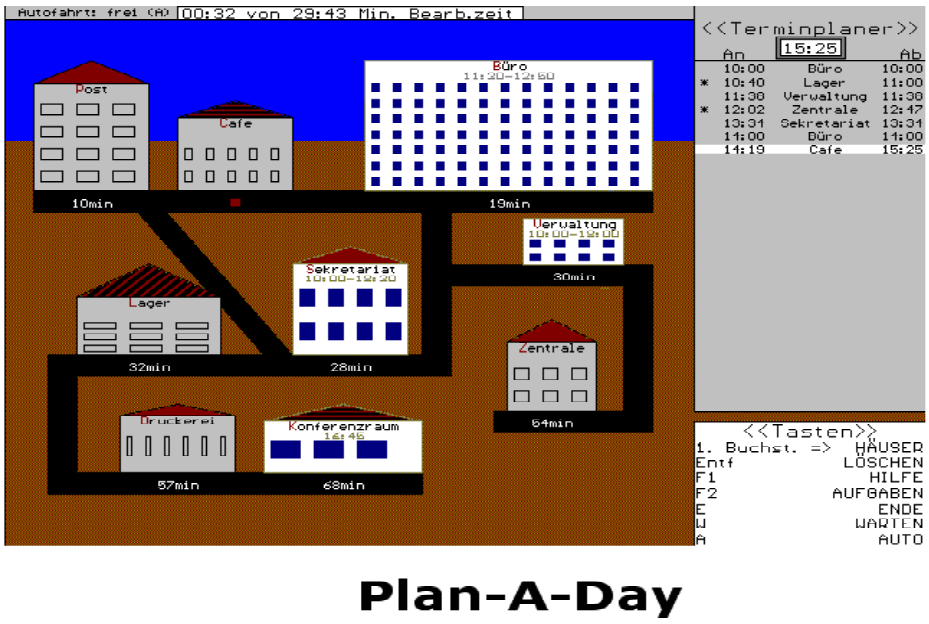


Figure. 2.2: PAD Interface.

The position of the little square shows that the participant is currently at the café. The locations on the map, which are colored white instead, of gray are locations at which a scheduled appointment hasn't been met yet. The times at which the appointments can be met are displayed in the roof of the respective houses. Below the locations, the distance from the participant's current location is given in minutes.

In the upper right part of the screen (headed "Terminplaner"), the current state of the system

current schedule, including the times of arrival and departure for the individual locations. It can be seen that, before the visit to the café, the participant has already been to the Storehouse, the Administration, the Central Office, the Secretary and his/ her own office – so s/he is lucky that the current appointment is taking place at a location where caffeine supply is imminent.

Every time the participant moves to another location and adds this location to the schedule, this moves and the times associated with it (arrival and departure) will be added to the "Terminplaner" (schedule). If the participant deletes a move, it is also deleted from the schedule, so abandoned schedules are not retained on the screen but have to be stored in memory.

In the lower right part of the screen, the functions of several keys are listed: The participants are, again, reminded that they can move to a location by pressing the key bearing its first letter, and which keys they have to hit to take another look at this days appointments, to obtain general help about the system, to delete a move, to declare the schedule finished and, finally, to take the car for the next move.

The participants receive direct feedback about their scheduling behavior *only if* they have made an "impossible" move. "Impossibility" is restricted to the case that the participant arrives at a location after the last possible point of time to meet this appointment has expired. The participants receive no general feedback about the quality of their schedules, nor are they forced to do every subtask within a day. Thus, apart from the possibility-constraint, the resulting schedule is up to the participant.

## 2. 4 How human scheduling is assessed by PAD

### *2.4.1 Performance*

As already hinted at before, PAD provides several options to evaluate participants scheduling behavior, which shall be described here briefly.

For the quality of the solutions (i.e. the final schedules participants come up with), a weighted and a transformed score are computed. The *weighted score* is the sum of appointments met, weighted by the priorities associated with these appointments (no priority mentioned/unimportant = 1; important = 3; very important = 8). The *transformed score* is created to take into account the fact that it is possible to achieve a considerably high score even without paying attention to the appointments' priority. It is computed thus: the

subtracted from the actual score, and two divides the remainder. Thus, the higher scores are transformed to take on values between 0 and 10.

During the course of planning, many participants create schedules that are "better" (i.e. yield a better score) than their final schedule. As it would be unfair to ignore this, a weighted and transformed Score is computed for the best schedule found by the participant as well. These scores are called the weighted and transformed "Max Score", as opposed to the analogously computed weighted and transformed "End Score".

### *2.4.2 Process: Heuristics and more (a peep into the future)*

Funke and Krüger (1995, 1995) repeatedly emphasize the importance of analyzing the scheduling process (as a whole) instead of only assessing the results of that process. Thus, they state that "the (Log-files) are of special relevance for future scientific investigations" (Funke & Krüger, 1993, p 9). They suggest a number of interesting possibilities for an analysis of the Log-files. The proposal to systematically investigate the extent of schedule construction/modification and looking ahead has already been mentioned in the introduction to this thesis and needs no further highlighting here. However, it is interesting to note here that Funke & Krüger (1995) already make an intuitively appealing distinction between a spontaneous "restart" in the scheduling process (a schedule is discarded completely and another is developed) and a local modification/ optimization of an already existing schedule. They also offer a preliminary explanation of the first kind of behavior as an example of "ad-hocismus" sensu Dörner (1989), while the second approach receives the slightly more favorable classification as specimen of evolutionary Optimization (as in DNA Computing, see e.g. Pisanti (1997) for an overview). While the question of the relative utility of the two manners of schedule modification is certainly open to discussion, the distinction itself is inspiring, which is of course the reason I chose to pursue it in this thesis.

The two proposals mentioned above were the two suggestions made by Funke & Krüger (1995) that bear the most relevance to, and are dealt with, in this thesis. There are many roads that are yet non-pursued in the jungle of human planning and scheduling. However, an interesting tool to analyze the Log-files collected by PAD is already implemented in the system: It is possible to measure which heuristics are likely to have influenced the *choice of the next appointment* during the course of planning.

While Funke and Krüger (1993, 1995) specify nine plausible heuristics that may influence scheduling behavior in the domain of PAD, only five of these have been included in the PAD-

- Meet the closest appointment first (minimize distances)
- Use the Car for a long distance to maximize the resulting advantage
- Meet the (very) important appointments first (mind priority)
- Meet the most urgent appointments first (mind urgency)
- Avoid too much waiting time

It is easy to see that the use of heuristics (1) – (3) can be deduced after each move simply by inspecting the appointment that has actually been picked. This is the case, because each of the last three heuristics exploits one of the criteria that is associated with all appointments (in the description of the appointments for one day, the earliest and latest possible time to meet them are mentioned, as well as their priority). Thus, at each moment during the course of scheduling, one or more appointments can be found achieve the score "1" given the application of one of the first three heuristics (i.e. the most urgent appointment, the most important appointment, etc.). The other two heuristics use internal system information, but they can be implemented in any computer program that represents the distances between the locations in a suitable data-structure.

The results of the analysis of heuristic application are summarized in the following manner: the average ranking (computed using all choices made during the course of scheduling) for each heuristic is compared with a value that would be associated with that heuristic if the choices of the participants were completely random. Because the choice of an appointment that ranks highest ("1") with respect to a heuristic is taken as evidence for the application of a heuristic, a low average value for a heuristic indicates a higher frequency of its use.

The option to analyze the Log-files with regard to specific heuristics in order to assess their overall application is both interesting and neat. The approach to implement the analysis directly in the system, so that it can be performed automatically, combined with the very clear and specific description of the heuristics themselves, avoids many of the dangers of verbal/ written protocol analysis (e.g. loss of information, low inter-rater reliability, ambiguity). (However, many of these dangers will be encountered again in chapter 5.)
It is also possible to easily test specific hypotheses about the predominance of a heuristic, given a specially constructed PAD task, or some other, more sophisticated, experimental intervention.

An especially interesting application of the heuristic facility lies in the field of Cognitive Modeling. To accurately fit and predict the preference for specific heuristics, and the overall distribution of scores for the five implemented heuristics, is an extraordinarily sublime test for any cognitive model. An example of this approach can be found in the Diploma Thesis of Wolfram Schenck (2001), in which a connectionist model of human scheduling in the PAD domain is presented. Although this model has some (minor) problems[8], it amazingly fits not only measures of human performance in the PAD task, but also predicts various measures of the scheduling process, e.g. the distribution of operator use (=moves to the locations) and the proportions of being on time or too late. Schenck's model also predicts the use of the individual heuristics. While this *may* be an artifact of the ambiguity resulting from the simplicity of the heuristics (see paragraph below), it nevertheless renders considerable support to his model. Furthermore, it is easy to think of research objectives that involve the development of Cognitive models of (e.g.) the dominance of specific heuristics under different conditions (e.g. low versus high time pressure). The heuristic-analysis facility implemented in PAD makes it easy to test the predictive value of such models.

However, in spite of all the advantages that come with the present analysis of heuristics, the kind of heuristics that were described above may be insufficient to describe human scheduling, because they are both too simple and too specific (they only take one criterion of the appointments into account). Huchler (1999, p. 74) has already hinted at the fact that adhering to only one heuristic is not sufficient so solve a PAD task. She also states that the heuristics are not mutually exclusive, and uses the example of the heuristic "to meet as many appointments as possible" (pp. 91-92). This heuristic clearly requires adherence to other "sub"– heuristics as well, e.g. to the heuristic to minimize the distances and the waiting time.
Another problem with such "simple heuristics" concerns the analysis of empirical data with regard to the application of these heuristics. The following problem arises: one criterion taken individually, be it the distance, the start-time, or another, allows no unambiguous ranking between the appointments, if the other criteria aren't taken into account as well. As long as this information is neglected, an appointment can rank "highest" according to two different heuristics (=criteria), and the same heuristic can "favor" two different appointments. This causes ambiguity in the automatic analysis of the heuristics, and makes it difficult to draw

[8] ...the pointing out of which is not the purpose of this thesis, as dissecting a fellow students work not only shows

definite conclusions about the reasons that were really determining participants' choices of the next appointments, let alone make predictions about them.

For now, however, it should be stated that, despite these problems, the heuristic-analysis option that is realized in PAD is a promising step in the most interesting direction of human scheduling research. Perhaps this thesis can serve to provide some inspiration on how to enhance and extend this analytical method.

## 2.5 The Structure of the Log-files, some useful terminology and list of abbreviations

This last section of chapter 2 will be devoted to the introduction of the terminology that will be used in the remainder of the thesis to describe human scheduling behavior in the PAD World. This terminology is not based on any definitions already made in the literature on scheduling, and neither do I have the intention to propose it as some kind of standard. The terms I chose were intuitively plausible to me, and I hope this applies to the reader as well. Their purpose lies in making the explanations and discussion that follow in the subsequent chapter as clear and evident as possible.
To provide the reader not only with terminology, but also with a clear picture of what these terms designate, I will introduce this terminology using an (imaginary) Log-file as an example. The terms that will be relevant throughout the remainder of this thesis are printed in ***bold Italics.***

Consider the following plausible excerpt from a Log-file[9]:

- Move to the Conference
- Delete the move to the Conference
- Move to the Storehouse
- Move to the Café
- Move to the Secretary
- Delete the move to the Secretary
- Move to the Administration
- Move to the Secretary
- Delete the move to the Secretary

- Delete the move to the Administration
- Delete the move to the café
- Delete the move to the Storehouse
- Move to the Administration (...)

*Figure 2.3: Imaginary Log-file, "raw" format. This format very closely approximates an english translation of the German original, the format being a little neater.*

It is possible to gradually transform such a Log-file in a Lisp-like List-structure, which holds almost the same information as the "raw" file, with the additional benefits of making some details of the process more obvious and easy to detect.
Prior to the analyses described in chapter 4, all empirical data were transformed into this Lisp-compatible format, as the analytic procedures themselves were programmed in Lisp.

*Table 2.1: Transformation of a PAD Log-File into a Lisp-like List. Explanation is given in the text.*

| *Intermediate "abbreviated" version of the Log-file* | *Lisp-like List-structure* |
|---|---|
| Conference (delete) | (Conference) |
| Storehouse | (Storehouse Café Secretary) |
| Cafe | |
| Secretary (delete) | (Storehouse Café Administration Secretary) |
| Administration | |
| Secretary | |
| (delete delete delete delete) | |
| Administration | (Administration) |

Several things about this transformation are notable, the first being that the deletions of moves are no longer explicitly mentioned in the Lisp-structure. Instead, the following information can be drawn from the latter without so much as a second glance:
First, there is the number of ***modifications*** made to a schedule. This is simply equivalent to the number of new lists generated. In the example, there are four lists, which means that there have been four modifications to the schedule. Note that a new list is created only if an element

create a new list. This is relevant in the fourth chapter, when the issue of modifications of a schedule will be examined in more detail. The schedules that are being modified, i.e. all schedules apart from the last schedule, which is the ***solution***, are simply called ***partial schedules*** (no need to be overtly creative here). The complete schedule (partial schedules and solution) will be referred to as, indeed, ***complete schedule***.

In the example, we also see two instances of a special kind of modification: A ***complete restart, or switch*** (to put it a little less formal). This means that a partial schedule that starts with an appointment is abandoned and another appointment is placed at the start of the schedule. A complete restart takes place after the move to the Conference and after the two partial schedules that start with "Storehouse".

It is now time to introduce the concept of ***modification-extent***. The term "modification-extent" describes an appointment. It is used to describe how many modifications to partial schedules starting with that specific appointment exist within a given course of scheduling. In our example, the modification extent of the appointment at the Storehouse is 2; the modification extent of the appointments at the Conference and at the Administration is 1. This will be relevant with regard to the questions about local optimization vs. discarding a schedule: Obviously, the greater the modification-extent of an appointment is, the more local modifications/ optimization-attempts are associated with it. The interpretation of this measure *must*, however, be qualified thus: In the case that the modifications of a partial schedule starting with a specific appointment are discarded in favor of another appointment, but resumed later, it must be differentiated between the ***overall modification-extent*** and the ***longest modification phase.*** The latter designates the longest uninterrupted modification-extent of an appointment during the scheduling process of a single participant (algorithm), the former is the sum of all modification-phases of an appointment during this scheduling process. This is important to distinguish between continuous work on a partial schedule and frequent discarding and resuming of schedules. Specific ideas about the behavior underlying the possible combinations of a long/short ***longest modification phase*** and a small/ large ***overall modification extent*** will be expressed concisely (and used for data-analysis and interpretation) in chapters 3 and 4.

Three other terms are important. Firstly, there is the ***modification-length***, that is, you guessed it, the length of a modification. The average modification-length for a participant can be computed, as well as the average modification-length for an appointment and a group of participants. In the Log-file above, the average modification-length of the appointments at the

the Storehouse is 3,5. The average modification-length of the "participant" is approximately 2,2. This measure can be important to test assumptions about the length of specific modifications, as well as differences in the average modification-length between groups of participants.

Secondly, there is the ***variety***, which designates how many different ***modifications*** there are in a course of scheduling, i.e. how many different appointments are placed at the beginning of a schedule during the course of scheduling. This measure is important to qualify the number of ***restarts***. Consider the example log-file again. In this protocol, we find two restarts and a variety of three (three different modifications): One modification of the Conference and the Administration each, and two of the Storehouse. This indicates many restarts, as well as a high variety. It is, however, also imaginable that a participant produces many restarts, but little variety, e.g. by switching between two appointments, which indicates different scheduling behavior. The relevance of this distinction should be obvious. Chapter 4 will address the question which kind of scheduling behavior is actually exhibited by humans, and how these measures (variety and restarts) correlate with the number of modifications.

Thirdly, the ***possibility*** of the modifications is of course interesting. This measure indicates if the schedulers have arrived too late at the latest appointment of a modification, or if they were in time. In the latter case, the possibility, is t (true) and in the former case (of course) nil. As the protocol above is a product of fantasy, it is not possible to exemplify this notion, however, the interpretation of this measure can be explained thus: a ***low number of possibilities*** can indicate either insufficient look-ahead or sloppy pre-calculation. A ***high number of possibilitie***s in a course of scheduling is a somewhat ambient phenomenon: If it correlates with a high number of modifications, it may indicate unnecessarily many modifications or restarts, when it correlates with a low number of modifications, it could indicate "good" look-ahead (i.e. correct calculations). These musings are beyond the scope of this chapter, and will be elaborated in the two subsequent chapters in more detail.

### *2.5.1 A small problem*

There is one problem (or rather: peculiarity) about the Lisp-like format of the log files. Consider the following two modifications:

(Storehouse printing-Office cafe Conference)

(Storehouse printing-Office Conference Secretary)

It is not possible to determine if the person who produced these two modifications has only deleted the last two appointments in the first modification (the cafe and the Conference), and has inserted the Conference and Secretary afterwards, or if (s)he has deleted the complete schedule and re-entered it ("Storehouse, printing-Office") before adding the two last appointments. This could constitute a problem, because the latter would formally be a restart, while the former is a local modification. However, I hold the view that as long as I chose to pursue a particular path of scheduling (as, in this case, to start my schedule with the appointments Storehouse and Printing Office), it is secondary whether I re-enter that schedule or whether I maintain it and modify its latter part. The critical fact is the maintenance of this schedule.

I also want to add that the PAD Interface makes it much more plausible to maintain the beginning of a schedule (instead of deleting and re-entering the complete schedule only because I want to change something at the end). The build-in "Terminplaner" makes it easy to maintain the beginning of a schedule and only make changes where it is necessary, and my observations during the studies I carried out for this thesis confirms this.

The assumption that participants maintain the beginning of a schedule and do not re-enter it every time they modify it is further supported by some data reported by Wolfram Schenck (2001). He reports the average number of successive deletions participants[10] exhibited in PAD 4 and PAD 5. This number is 2.2 for both PAD 4 and PAD 5 (p.71). Additionally, in my own analysis of the (same) data I found that the average length of partial schedules in both PAD 4 and PAD 5 is 4. That makes it extremely unlikely that participants delete the complete schedule every time.

Schenck (2001) offers additional evidence for this absence of complete deletions. According to his analysis (pp.71 – 73), the number of complete deletions of a schedule is only approximately 3, for both PAD 4 and PAD 5.

However, I admit that the problem described in this paragraph introduces some ambiguity into the subsequent data-analysis. This was one of the reasons to introduce the measure of variety to qualify the measure of the restarts. The subsequent data-analysis will thus rely for the most part on those two measures, which are, in combination, not ambiguous.

### *2.5.2 List of abbreviations of the appointments*

The following table holds an overview of all appointments that have to be scheduled in PAD 4 and PAD 5. Although these two PAD tasks partly involve identical appointments (middle column), the times at which these appointments take place is not the same in PAD 4 and PAD 5.

*Table 2.2: Overview of the appointments in PAD 4 and PAD 5, with abbreviations.*

| *Appointments in PAD 4* | *Appointments in PAD4 and PAD5* | *Appointments in PAD5* |
|---|---|---|
| **Secretary**: **S** | **Printing Office**: **PO** | **C**afe: **C** |
| **Storehouse** : **St** | **Conference**: **CO** | **Central Office**: **Cent** |
| | **Administration**: **AD** | **Office**: **O** |

# 3 Theoretical Musings

This chapter is devoted to a more detailed analysis of PAD. I attempt a theoretical classification of the behavior PAD elicits.

I will first analyze PAD as a Constraint Satisfaction Problem and show how the criteria of its appointments will influence the difficulty of a PAD task.

Afterwards, I will compare PAD to the paradigm of classical planning. While the PAD scenario meets many constraints that prevail in this paradigm, the task itself is closer to a problem-solving task than to planning per se.

I will then examine three theoretical and computational approaches that claim to be both cognitively plausible and efficient in dealing with particularly complex tasks: "Adaptive Planning" (Altermann, 1988) which derives from Schank & Abelson 's (1977) Script theory, "Opportunistic Planning" described by Hayes-Roth and Hayes-Roth (1979) and Anderson' s (1987) concept of skill acquisition -based on his ACT* (1983) theory- which states that domain-specific skills are the result of weak problem solving methods that operate on general declarative knowledge people have about a task or domain.

I will use all these approaches to guide further analysis of behavior in the PAD world. Specifically, I will investigate the role of declarative and procedural learning in PAD. Declarative learning in the PAD domain is conceptualized as the accumulation of experience, which is achieved by exploration, i.e. trying out partial schedules. Procedural learning in the domain of PAD concerns the speedup of the mental arithmetic that is applied in the selection of the next appointment. I will show how these two kinds of learning can work together to produce good scheduling, as the development of the latter skill enhances the quality of the exploration.

## 3.1 The Complexity of PAD and its non-existent consequences

To show that a problem is NP complete, the usual strategy is to show that another problem, the NP-Completeness of which is already known, can be reduced to the problem in question (see, e.g., Sipser, 1997, for the general procedure, and Garey & Johnson, 1979, for a collection of NP complete problems with the respective proofs). In the case of PAD, the classic Travelling Salesman Problem (TSP) offers itself. I will not give a formal (mathematical) proof here, but instead outline the main argument of the reduction, which is sufficient for the current purpose.

The TSP can be stated in the form of the following yes/no question: Given a map depicting various cities, which are connected by roads of variable length, is there a path that connects all cities and that is shorter than a fixed length "d"?

Any TSP can be changed into a PAD problem by using the following transformations: The cities are the appointments (which, for the sake of the argument, take zero time). The roads

are the distances between the appointments. The distance "d" is the time from 10 a.m until the latest possible time to do an appointment.
Of course, although this argument is "relatively straightforward", the resulting PAD-task is "a bit of a strange task, without any constraints and zero-time appointments" (both quotations courtesy of Niels Taatgen, personal conversation). As we have seen in the preceding chapter, the existence of constraints (i.e. "time windows") is an important defining characteristic of a PAD-task. The same holds true for the duration of the errands, which naturally has to be included in PAD to maintain its much-stressed realistic context. This shows that the consequences of PAD's NP-completeness only take effect in a highly constructed worst case.[11]

This is not only true for PAD, however. Most instances of NP-complete problems come with constrains that make it easier for machines or humans to cope with them. This "coping" is usually referred to as "Constraint Satisfaction Search" (CSS), and the respective problem is called a "Constraint Satisfaction Problem" (CSP) (e.g. Russell & Norvig, 1995, p. 83, p.104). Let me explain the concept of a CSP using PAD as an example[12].

A CSP is usually stated as a set of variables, a set of possible values, and a set of constraints that the values have to obey. The problem solver must assign a (set of) value(s) to each variable in such a way that no constraint is violated.
Just exactly in which way one wants to map a particular problem onto the CSP formalism is always a bit of an arbitrary matter. In the course of writing this thesis I have devised multiple definitions of PAD as a CSP, and found the one that follows the most pleasing. However, this mapping is certainly not the only one that is possible.
In PAD, the set of variables contains the positions in the schedule. If a PAD task contains 6 appointments (including the car option), the variables are positions 1 to 6. The values are the appointments. The constraints are the time-windows (i.e. the "space" between earliest and latest time) of the appointments.

---

[11] Of course, it is exactly this worst case that is crucial for the classification of a problem (or task) in terms of its complexity (Sipser, 1997).

[12] For purposes of readability I have decided to give an informal explanation in this text. However, this explanation was created in exact analogy to Ginsberg (1993), where the interested reader can find the formal

This last point carries the implication that the quality of the constraints is crucial to the easy or hard nature of the particular instance of a problem. PAD is easy if the time to do the appointments is constrained in such a way as to create a linear ordering among them. In that case, the appointments can simply be met one after the other. PAD becomes harder the more intersections exist between the time windows of the appointments, because in that case, it is harder to choose among them, and the risk to choose the wrong appointment next is greater. If the time window for all appointments is identical, the problem is hardest. At least, in the case of PAD, the additional information about the duration of the appointments and the distances between them can offer some more decision guidelines (it can be used as a substitute constraint, in case the time information is not sufficient). However, if this additional information does not support an unambiguous choice either, an irresolvable conflict arises, and one or more appointments cannot be met.

Funke & Krüger (1997, cited in Huchler, 1999, p. 82) have also commented on the difficulty of PAD tasks as a function of the intersection between the appointments. They claim that the difficulty of a PAD task is highest when the time windows of the appointments are largely congruent, but the appointments themselves are not completely mutually exclusive.
This difficulty results from the fact that participants now have to search for the right solution actively.
Niels Taatgen (personal conversation) has also pointed me to the fact that CSP are hardest with an intermediate number of constraints, because the extreme cases of no constraints and many constraints are trivial. In PAD, the notion of an “intermediate number of constraints” corresponds to what could be called the “intermediate discriminating value” of the constraints.

There exist a number of heuristics for Constraint Satisfaction Problems that enhance performance even in hard cases (Russell & Norvig, 1995, p.104). These heuristics use the methods of “forward checking” and “backtracking”. The latter method analyzes the search that has occurred until the current moment in order to avoid repeating states, and to keep track of dead ends. The former method looks into the future in order to avoid states in which the problems become unsolvable. I will discuss their applicability to PAD in section 3.6.x, which examines possible weak methods for PAD.

In the light of the preceding discussion it is no surprise that the property that is associated

their size (i.e., in PAD, with the number of appointments) does not appear in the empirical data that were obtained during studies that used PAD. In a study described in Funke & Krüger (1995, p.115), one group of participants had to solve PAD tasks 4 and 5, and another 13 and 14. The two latter tasks contain nine appointments each, The two former five and six appointments. Despite this considerable difference in size, participants take on average the same time for the "smaller" and "larger" tasks (616 and 699 sec for PAD 4/5, respectively and 755 and 568 sec for PAD 13/14 respectively). A similar pattern was observed for the number of "operations" (i.e. movements to locations, car-use and deletions). Participants that had to solve PAD 4 and PAD 5 used (on average) 16 and 17 actions, respectively. Participants who solved PAD 13 and PAD 14 used 19 and 23 actions, respectively. While these findings are moderated by the fact that there are multiple solutions to PAD 13 and 14 and only one for PAD 4 and 5, the moderate difference between these two groups of tasks nevertheless speaks a clear language. Moreover, PAD 4 and 5 also differ with respect to their size,[13]but hardly with respect to the time and actions needed to solve them.

In the data I collected in the study described in chapter 4, a similar pattern emerges. Participants took, on average 8 min. to solve PAD 4 and 9 min. to solve PAD 5. Moreover, the total number of modifications that were produced by participants while they worked at PAD4 was 325, and only 340 during the work on PAD 5. As 43 people participated in that study, that's less than one modification more (on average) per participant.
Other measures such as the number of deletions, and, consequently, the ratio of deletions to actions, remain almost uncannily stable between the two tasks (the total number of deletions is 655 is PAD 4 and 692 in PAD 5, the average ratio of deletions and actions is 0.32 in PAD 4 and 0.29 in PAD 5).

However, as mentioned before, all of this is not really surprising, as the NP-completeness argument only holds for the worst case anyway. Moreover, as Hertzberg (1995) states, the fact that humans do not show the exponential rise in required time, can either indicate that the

[13] Funke & Krüger (1993, p. 6) show a way to compute the set of "rational solutions" for each PAD task. This is equal to the combinations of all tasks, excluding visits to locations where no appointments are scheduled and visits that place a later appointment before an earlier (more constrained) one. The number of rational solutions for PAD 4 is 101, and for PAD 5, it is 388. This difference makes the "stability" of human scheduling behavior even more compelling.

"worst case" hasn't been met by a particular instance of the task, or that the underlying mechanism in solving the task is different from the "classical" notion of planning (for more on that notion, see the next section). We have already seen that the former is almost always the case with PAD, so it's time to explore the latter.

### 3.2 A brief excursion to planning in AI

A brief comment must be made to justify my selection of theories to be examined in this chapter, which could perhaps be called representative (although even that point is open to discussion) but certainly by no means complete.

To explain this choice, I have to concern myself a little[14] with the ideas of "classical planning", as it has dominated research in AI for a long time. Let me first review what is meant by the term "classical planning".

Planning as such is often described in AI as finding a sequence of actions that will yield a specific goal. Russell & Norvig (1995) summarize: "Planning agents use look-ahead to come up with actions that will contribute to goal achievement."(p.362). They are similar to problem solving agents, but not entirely identical. As I will (in accordance with Schenck, 2001) argue later in this chapter that PAD is closer to (general) problem solving than to (classical) planning, it is worth to briefly highlight these differences here (taken from Russell & Norvig, 1995, pp. 338 – 341).

- A more open representation of states, goals and operators in form of sentences enables planning agents to detect relation between states and actions
- The planner can insert actions in to the plan when they are needed, while the problem solving agent works with an incremental sequence starting with an initial state and proceeding in one direction
- Planners exploit the fact that most parts of the world are independent of another by creating partial sub-plans that can be carried out separately and combined in the end; this is a "divide and conquer" – strategy.

[14] For an extensive (and funny) overview on the field of planning in AI, which covers more recent as well as

To enable artificial systems, i.e. algorithms and computing machines, to perform this task, several constraints had to be established. These constraints constitute the frame of classical planning. The ten most important of these constraints are (translated by Schenck, 2001; originally from Hertzberg, 1995):

- There exists only one planner (planning actor)
- It is possible to represent the relevant parts of the world in states; these states are complete snapshots of the world
- State transformations by planned actions are the only form in which time is represented
- Planning and plan execution are carried out one after another
- Complete information about the facts within the "world" are available during planning as well as during plan execution
- The effects of an action are deterministic and context-free. That means, they are identical for every state in which the action is executable.
- During plan execution the world is only changed by the actions of the actor, who is guided by the plan
- The objectives of the resulting plan are explicitly stated; they are consistent and can be achieved by known actions.

It is easy to see that some of these constraints are violated in "real life" planning or scheduling, e.g. the completeness of information, the non-interruptibility of the planner, and the infinite amount of time. This problem has already been mentioned in the introduction.
This lack of (cognitive) plausibility does not constitute a problem in itself, as it is not the objective of AI to accurately model human behavior –this is the aim of cognitive modeling. AI uses specific features of human thought in order to develop algorithms that can solve a wide range of task efficiently. Cognitive Modeling imitates, and AI creates, which is perfectly legitimate[15]. However, "classical" planners also face problems within the domain of AI. These problems usually stem form the intrinsically hard nature of some problems, as, e.g., PAD, which causes an inflexible "classical" planner to use a lot of computation time.
Interestingly, some of these problems have been tackled by introducing mechanisms that are, implicitly or explicitly, more cognitively plausible. For example, the planner STRIPS (Fikes & Nilsson, 1971) is based on means-end analysis and adheres to the principles of classical

[15] Following this line of reasoning, Newell & Simon's (1972) work in the general problem solver (GPS) must be

planning. While STRIPS provides us with a neat paradigm to code operators and states for a given problem[16], it faced some problems that, only one year later, resulted in the inclusion of Macro-operators (Fikes & Nilsson, 1972). These Macro-operators test whether abstract plans can apply to a new situation; i.e. a plan need not be created from scratch anytime a new problem arises. This resembles a rudimentary memory system. Other planners that behave more human-like (a collection of them can be found in Akyürek, 1992) employ analogical reasoning from examples, also a familiar feature of human problem solving (Anderson, 1983; Anderson, 1987; Anderson, 1986; Anderson & Lebiere, 1998).

Other, more specific, improvements from the already mentioned field of Constraint Satisfaction Search are also aimed at using memory more efficiently by establishing sophisticated backtracking strategies that retain successful parts of the solution to the problem at hand, modify only faulty parts, avoid redundant search, and favor local instead of global modifications (Ginsberg, 1993). A related approach is the analysis of "dead ends" that have occurred in the problem-solving process, in order to avoid the same mistakes, in combination with more or less sophisticated look-ahead methods (Dechter & Frost, 2002). These ideas implement in effect a rudimentary learning mechanism. I will take them up again in discussing Anderson's (1983, 1987) theory of skill acquisition.

The preceding paragraphs have been quite critical of classical planning, and give the impression of portraying the "good influence" of psychologically plausible constructs like episodic memory or learning on the field of AI. It is, however, not the intention of this thesis to refute one specific theory, or school of research. That would be trivial indeed, especially given Hertzberg' s (1995) statement, already cited in the introduction, that "there is no thing as planning as such". Instead, it is worthwhile to ask: "To what degree are these particular ideas relevant for PAD?" This shall guide further analysis.

So, to what degree are the ideas of classical planning relevant for PAD?

Schenck (2001) interestingly points out that some of the constraints of classical planning are met in the PAD world, e. g., there *is* only one planning actor, the PAD world *can* be represented by states (of the "Terminplaner") that are in themselves complete. There are no

[16] In STRIPS, states and operators are coded in terms of first order logic. The description of states contains the difference to former states, and the description of operators contains the changes they can make to any given combination (formula) of states. This is much more efficient than, e.g., an endless list of "if...then...else"

"hidden layers" or dynamics, which produce surprising outcomes: the constraint that during plan-execution the world can *only* be changed by the actor holds, too. Although the effect of a move to an appointment depends on the position of that appointment in the already existing schedule (i.e. I can be too late if I go to the Conference after the café but in time vice versa), the effect itself is predictable. Given the same context, it remains always the same, meeting the sixth constraint. Thus, PAD as a task can be classified in close proximity to classic problem-solving tasks that can be solved by classic means, as, e.g. means-end analysis. PAD is not a highly complex, dynamic and unpredictable real world scenario.

Schenck (2001) notices the following subtle distinction/ interaction between planning and problem solving in the domain of PAD. PAD requires participants to schedule appointments, i.e. find a sequence of operators, "and this is clearly a planning problem" (Schenck, 2001, p.28). However, the fact that participants can delete moves places PAD close to Problem-Solving in a more general sense, "where operators may be undone, and where the problem solving process may go back and forth to every known state in the problem space" (p. 28 – 29). The PAD Interface also clearly evokes the incremental construction of a sequence of operators (moves to appointments), starting from an initial state (Office). According to the definition of a planning agent (Russell & Norvig, 1995) given above, this rather calls for a simple problem-solving agent than one for planning.

The stages of plan-preparation and plan-execution are intermingled in PAD. This makes the process more vulnerable to disruptions (trial and error behavior, bottom–up planning), because "wrong" decisions have no direct harmful consequences.

A more severe "no return" scenario, in which time passes as in real life and cannot be recovered would probably produce a slightly different, presumably more deliberate, kind of behavior, and perhaps better plans as well. However, the value of PAD lies precisely in its flexibility, which enables the researcher to witness the search-process that ultimately leads to the complete schedule. By allowing for mistakes and modifications, PAD lends as much transparency to the flow of human thinking as can be obtained without the use of verbal-protocol analysis.

To sum it up, both the paradigm of classical planning and the paradigm of problem solving prevail in the PAD world.

However, the data mentioned at the beginning of this chapter (section 3.1), concerning the latencies and the number of actions in different PAD-tasks suggest that humans must have some method to avoid the dangers that have to be faced by classical planners. This was the reason to introduce some psychologically motivated theories of planning and scheduling. The former paragraphs on classical planning helps to justify my choice in this regard.

The three theoretical approaches I will now discuss are each prototypical of a specific element of cognitive plausibility that was introduced into classical planning with the objective to enhance its performance.
Firstly, the accounts of Altermann (1988) and Schank & Abelson (1977) use the concept of episodic memory, remindful of the early modifications to STRIPS.
Secondly, there is the approach of Opportunistic planning, which emphasizes the fact that planning can also occur in a "bottom-up"-fashion, i.e. a plan can be changed throughout its execution. This possibility arises out of PAD's conceptual proximity to problem solving and the reversibility of actions, as pointed out by Schenck (2001).
Thirdly, there is Anderson's (1983, 1987) theory of the learning of Cognitive Skills, which can be connected to the mechanisms of "sophisticated forward checking", which de facto implement procedural learning throughout a problem solving session. This similarity is not obvious yet; however, it will become more clear in the course of the section of this chapter that is devoted to Anderson's (1983, 1987) theory.

I have already reported Schenck's (2001) assessment of the relevance of classical planning for the PAD world. I will now attempt a similar assessment with regard to the three theories mentioned above.

## 3.3 Memory, Scripts and Adaptive Planning: The ideas of Schank, Abelson and Altermann

Hertzberg (1989) summarizes one often-heard critique of the concept of planning in AI in the following statement "No one plans the solution to every-day 'problems'!" (p. 214). This is, to a certain degree, true. It is hard to disagree with Hertzberg 's elaboration of his statement: "If I am at home and discover that I'm hungry, I don't sit down and make a plan that tells me how I, by minimizing the product of time and path-length, may enter a state in which the statement "I'm full" is TRUE" (p.214).

According to the notion of Adaptive Planning, what we are likely to do instead is retrieve an old plan which has worked in the same or a similar situation before (e.g., call the Pizza Service). If we are in a situation that diverges from the situation in which he plan was originally carried out, we modify the plan. If the situation is the same, we simply execute it again.

The plan is checked with regard to its appropriateness for the current situation step by step. If a divergence is found, the step is adapted, if no divergence is found, it is incorporated in the current plan.

The crucial element here is the interpretation of the situation ("situation matching"). Altermann (1988) postulates three possible basic differences between planning steps. These differences are interpreted with regard to the situation as such: different preconditions of the planning steps, different outcomes, and different goal specifications.

The new plan is obtained by means of "abstracting" from the old one. This means that specific parts of that old plan have to be removed and changed, while the basic relations between the steps remains intact. To retain these relations, it is necessary to postulate a kind of background knowledge about situation. In Altermann (1988) this background knowledge encompasses categorizational knowledge, which uses the ISA-relationship among concepts, partonomic knowledge, causal knowledge and role knowledge. Causal knowledge in turn contains five types of relations: purpose, reason, goal, precondition and outcome. This network of knowledge is combined to make sure that the adaptation of an old plan and the substitution of the planning steps are carried out correctly.

After the abstraction, a possible candidate from the same category (obtained by means of abstraction) as the rejected planning step is selected and tested according to its applicability. If this alternative can be accepted, the adaptation process continues with the next step. If it is rejected, another candidate is selected.

Altermann (1988) distinguishes his approach from the superficially similar idea of case based reasoning (Carbonell, 1981, 1983).

Carbonell has also used the idea to apply earlier plans to novel situations. In contrast to Altermann, however, he emphasized the lack of abstract knowledge in many novel situations, and instead proposed a use of old plans by means of analogy. He offered two different approaches to implement this notion. In the first (1981) he transformed an old problem into a new one, using means-ends analysis. In the second (1983), he proposes a mechanism called "derivational analogy" to map the decision making process of the old problem

According to Altermann (1988), the main differences between his and Carbonell' s work are the following:

Altermann (1988) assumes that the specific plan is used in order to create an appropriate one for the current situation, with the more abstract plans serving as "backup strategy" in case the specific plan is partially inappropriate. In contrast, Carbonell (1981, 1983) assumes that specific plans serve the role of "backup strategies" in case no abstract plan is available.

Furthermore, the process of "refitting" the old plans differs: For Altermann (1988), the process of situation matching is crucial, which depends on specific declarative knowledge about these self-same situations, while Carbonell (1981, 1983) employs more traditional weak methods like analogy and means-ends analysis.

This, the third difference, according to Altermann (1988), lies in the character and use of background knowledge. Carbonell' s "derivational history" (1983) contains a decision making process, while background knowledge sensu Altermann denotes "the relationships between the prestored plan and the other pieces of knowledge that are related to it" (p. 418).

Altermann (1988) states that "Adaptive Planning is in the spirit of recent work in artificial intelligence on modeling (!) human memory (e.g. Schank, 1982)" (p.418). This may be a good moment to briefly review the script theory by Schank & Abelson (1977).

Schank & Abelson (1977) focused on the understanding, rather than the construction of plans. They assume that human memory is build around episodes rather than being organized in an abstract semantic network. Two basic concepts in this understanding of human memory are the script and the scheme. The latter contains general knowledge that can be applied in specific situations, if they are exemplary of the scheme. The former denotes a stereotypical sequence of events which is likely to be required in a specific situation (the well known textbook-example of the restaurant script needs no elaboration here). The script is active in a "variabilized" form and can be instantiated according to the specific situation. However, Schank & Abelson (1977) emphasize that the scripts are relatively constrained: "A script is made of slots and requirements about what can fill these slots" (p. 41). This is remindful of Altermann' s (1988) abstraction mechanism.

Schank & Abelson (1977) state that scripts and schemes "(do not) provide the apparatus for handling totally new situations" (p.41). That carries the following consequence: A person can only understand a situation in which they have been before, or, more generally, which they have encountered before. This knowledge helps them to interpret things.

According to Schank & Abelson (1977) it is only in dealing with completely novel situations that humans recur to planning at all. Their definition of a plan is similar to the definitions from AI literature cited above. They conceptualize a plan as a sequence of actions that is aimed at reaching a goal (or multiple goals). Plans contain knowledge about relations between events and about actions that can connect events with each other. This is remindful of standard definitions of problem solving, and Schank & Abelson (1977) indeed classify the construction of a new plan as problem - solving, as opposed to the mere retrieval of the appropriate script.

As in Altermann (1988), this "background knowledge" is more abstract than the specific knowledge (old plans or scripts respectively), and is *only* evoked if none of the latter is available. Schank & Abelson view scripts as specific instantiations of plans. Both Altermann (1988) and Schank & Abelson (1977) seem to regard planning as a kind of "backup strategy", which has to be employed if the more convenient retrieval doesn't work, either for lack of previous knowledge, or because the previous knowledge is not appropriate anymore because the situation has changed.

### 3.4 Opportunistic Planning

The work of Hayes-Roth & Hayes-Roth (1979) on opportunistic planning is an early example of cognitive modeling, because the authors implement their model as a computer simulation, the "behavior" (i.e. output trace) of which they subsequently compare with human behavior. Although Hayes-Roth & Hayes-Roth (1979) also want to show the efficacy and functionality of opportunistic planning per se, the main objective of their work is the analysis and accurate modeling of human planning.

In order to assess human planning, Hayes-Roth & Hayes-Roth (1979) use the "A day' s errands" task (subsequently abbreviated ADE), which resembles PAD, as the name already suggests. As in PAD, participants who work with the ADE have to schedule various appointments for a day. Participants also work with a map that shows a fictitious city. There are some notable differences between the tasks, however. For example, the time-constraints in the ADE aren't as rigid as in PAD. For some appointments a duration and a latest possible time is mentioned, but not for all. There are no priorities mentioned, and, more importantly, the distances between the locations aren't given. Thus, participants do not obtain a feedback in the case of being too late, because "too late" is not defined formally. Another difference

PAD 5, Hayes-Roth & Hayes-Roth (1979) designed ADE in a way that was supposed to make it impossible to meet each of the (many) appointments.
These differences between the tasks are important for the subsequent evaluation of the relevance of Hayes-Roth & Hayes-Roth' s (1979) model of planning for scheduling in the PAD world and will be revisited later.

Hayes-Roth & Hayes-Roth (1979) assume that the structure underlying planning is organized as a blackboard, which is, in turn, divided into five planes. They are called "Meta-Plan", "Plan", "Abstraction", "Execution" and "Knowledge Base". Each of these planes contains various levels of abstraction, i.e. with regard to how close they are to the actual execution of a step in the planning process. For example, the highest level of abstraction on the Knowledge-Base-Plane is "errands", followed by "layout" and "neighbors" (i.e. errands that are close to each other), with "routes" (between the errands) being the least abstract level.
Planning is described as the result of various planning "specialists" communicating with each other on the blackboard. The "specialists" each implement possible steps of planning, e.g. a step to a specific location, or, on a more abstract level, the adherence to a specific criterion in selecting the next appointments. The "specialists" are independent of each other. They are implemented in the form of production rules that are divided in a condition and an action part.
The planning process proceeds in cycles. In each cycle, all specialists whose conditions are matched by the current state propose their actions to be incorporated into the plan[17]. The actions of the specialists are not coordinated systematically. Instead, the specialists behave opportunistically by indiscriminately offering themselves for use. One specialist is selected, and a new cycle begins. The planning process stops when a good plan (either according to an external criterion or to the planner) has been developed, or, alternatively, when failure cannot be denied any longer.
The decisions of the specialists are noted on the blackboard, and subsequent specialists match their conditions against these entries.
The specialists are associated with specific planes and levels of the blackboard, and they only have to take the entries already made on these specific places into account when they execute their actions.

[17] This idea has appeared in some more recent production system architectures, which claim to be inspired by neural parallelism, e.g. Soar (in which the production rules whose conditions are matched fire in parallel)

Hayes-Roth & Hayes-Roth (1979) emphasize the special features of their model:
Because all (matching) specialists from *all levels* are allowed to propose themselves in each cycle, their model can account for bottom up as well as top down processes in planning. An example for the interplay between these two could be the following situation: The participant has decided to focus on a specific area of the town, because he has discovered that many errands have to be performed in that area. Thus, heading there will enable him to do many errands in quick succession. Up until now, his planning has been strictly top-down: a general strategy has been established which is now carried out in practice.
However, the following situation can occur during the execution of the plan that has been created this way: The participant suddenly discovered that another location, which hadn't figured in the previous plan, is situated close to his current location (e.g. the cafe across the street). Spontaneously, he decides to go there and "take it in" on the way. After that, he can either resume the original plan or abandon it completely in favor of a new approach that has been triggered by the interruption.
This last bit of planning (the sudden realization: "Oh! I can do that errand too, while I'm on the way") is certainly a bottom-up driven process (a specific percept changes –perhaps! - the more abstract strategy).

Hayes-Roth & Hayes-Roth (1979) claim that their model is flexible enough to handle complex tasks, and, due to its opportunistic structure, avoids the situation of getting stuck. They also present an implementation of their model as Interlisp-Simulation and compare its output with the verbal protocol that was produced by a participant working on the ADE task. They conclude that the general fit between the planning process that is produced by the model and the planning process that can be deduced from the utterances of the participant is sufficient enough to confirm their assumptions about opportunistic planning.
They furthermore state that the relative amount of "spontaneous" bottom-up driven behavior and more deliberate top-down driven reasoning depends on the specific circumstances of the task, or on the demand of the real-life situation.

This last point offers itself (quite opportunistically) to initiate an assessment of the above-described theories to scheduling in the PAD world.

## 3.5 Adaptive and Opportunistic planning in the PAD world: An assessment

The theories of Adaptive[18] and Opportunistic Planning appear to be widely apart. Adaptive planning focuses on the organization of knowledge and its importance for coping with novel situations. Opportunistic Planning describes the phenomenon of bi-directional processing during plan-execution and uses it to explain interruptibility and erratic behavior in humans.
While the content and organization of episodic memory and background knowledge are absolutely essential for Adaptive Planning, it is featured only slightly mysteriously in Hayes-Roth & Hayes-Roth' s (1979) description of Opportunistic Planning (memory is featured as entries on the blackboard, left there by previously employed specialists). On the other hand, while the interruption of the plan execution is part of the opportunistic model, it is not mentioned at all in Altermann' s (1988) account.

What the two models have in common, however, is a slightly negative view of planning. While Hayes-Roth & Hayes-Roth (1979) repeatedly stress the interruptibility of any, even good plans, and the plan itself as the quite random product of the chaotic competitions of unconnected demons (that have to be coordinated by the Homunculus of the central executive), Altermann (1988) and Schank & Abelson (1977) view the generation of new plans as a second-best strategy that only applies if retrieval (from memory) fails.

Both theories offer interesting ideas for a deeper task analysis of PAD. Let me start with Opportunistic Planning.

### *3.5.1 Opportunistic Planning in PAD*

It has already been pointed out that, due to the design of PAD, the stages of plan execution and planning itself are intermingled in the PAD world. This makes the process vulnerable to interruptions as they are reported in the work of Hayes-Roth & Hayes-Roth (1979).
This becomes even clearer when we (re-) consider the design of the PAD interface. The map-like arrangement of the locations makes it plausible that participants discover "all of a sudden" that they are close to an appointment that was not part of the original plan, but is so conveniently situated that it can be done anyway. The PAD interface also enables the

[18] For the sake of verbal elegance I will use the term "Adaptive Planning" to denote both Altermann' s (1988)

participants to easily include these appointments in their schedules, either by just attaching it to the end or by modifying the already existing schedule.
This, however, already points to a divergence between the PAD world and the ADE world.

Hayes-Roth & Hayes-Roth (1979) report multiple examples in which their participant completely abandons a strategy, seemingly forgetting about it, and continues his plan "elsewhere", i.e. at another location or level of abstraction. While these shifts of reasoning are certainly in accordance with the notions of Opportunistic Planning, they are also supported by the special situation the participant was placed in. The participant did not have to carry out his plan, but instead had to describe what he would do with the errands he was assigned, looking on the map. He was not given a feedback on the quality of his plan, either. This resulted in him producing a plan that, amazingly, enabled him to do all errands on the list (which was constructed with the purpose of evoking an errand-overload!). Finally, he also wasn't allowed any means to remember his partial plans during planning.
In PAD, however, participants obtain immediate feedback in case they are too late. They can also inspect their current schedule at any time.
It is obvious that the specific setting of the ADE task in Hayes-Roth & Hayes-Roth' s (1979) study is more likely to evoke the "chaotic" behavior the authors describe as opportunistic. This behavior is probably connected to the fact that the participant wasn't able to correctly remember the partial plans he had already formed, and the lack of constraints he was faced with. Assuming that the Hayes-Rothian specialists do indeed exist, they were certainly given full play in their study.
On the other hand, life in the PAD world is much more constrained. This should result in a smaller amount of truly "opportunistic" behavior, due to the fact that the time constraints for each appointment, and all distances between them, are continuously available to the participants. Furthermore, due to the presence of the "Terminplaner", modifications can be made much more precisely (plans need not be abandoned completely), and the consequences of modification are immediately obvious (as the modifications have to be entered into the computer, which also elicits direct feedback).

Although the above paragraph shows some critical aspects of Hayes-Roth & Hayes-Roth' s (1979) work, they do not necessarily imply that their notion of Opportunistic Planning is completely mistaken. On the contrary, the close connection between planning and plan

spontaneous modifications. However, chances are that they will be a lot less "violent" than in those reported by Hayes-Roth & Hayes-Roth (1979) (who, after all, have already stated that the amount of bottom-up planning is likely to vary with the characteristics of the situation at hand).

Let us now leave these slightly fuzzy theoretical speculations behind in favor of more specific speculations. Given the characteristics of the Log-files described in chapter 2, what patterns could be indicators for the presence of opportunistic planning in the PAD world?
This question is hard to answer. Nevertheless, the following attempt can be made.

In principle, every modification to a schedule can be the result of an opportunistic demon piping in with an alternative move. However, modifications can also be the result of sloppy calculations and subsequent "impossibility" -feedback of the system.
Similarly, restarts can be the result of spontaneous opportunistic intrusion, but it can also occur after a series of systematic, yet unsuccessful, modifications to a schedule.

I therefore tentatively propose that the following patterns in the scheduling process could be called "opportunistic":

- Many modifications can be a sign for opportunistic planning, *especially if* they occur "spontaneously" (i.e. they are not prompted by the system).
- Many restarts can be a sign of opportunistic processes, *especially if* they occur with only few "local" modifications (to the end of schedules) in between (a short "longest-modification-phase")
- Another indicator for processes of opportunistic planning could be a high variety of the schedules.

Yet another indicator could be the length of the partial schedules: If a participant consistently produces short schedules, this could indicate opportunistic processes, or at least a certain readiness to modify the schedule quickly. This pattern of behavior would be consistent with a high variety in modifications and many spontaneous restarts.

All of this is, however, still quite speculative. It would certainly not be legitimate to analyze empirical data, search for the features sketched in the paragraph above, and conclude (if they are found): "Hayes-Roth & Hayes-Roth were right after all".
This would be unwarranted for various reasons.
Firstly, the inter-individual differences in complex tasks (and PAD is complex) are usually high, so it is unlikely that one pattern of behavior will be exhibited by all participants.
Secondly, empirical data only record overt behavior and not the underlying processes. We can therefore only analyze patterns and describe patterns. It is possible to describe an "opportunistic pattern", which refers to a pattern of behavior that would be consistent with the notion of Opportunistic Planning – but could also be the result of different processes, as we shall see in the next paragraph.
However, without verbal protocols, we may not state that the processes underlying these patterns are indeed identical to the processes postulated by the adepts of Opportunistic or Adaptive (or another optional attribute) Planning.

This should be kept in mind throughout the remainder of this chapter, as well as in the next chapter, which features an analysis of the patterns of modifications found in human data, and, naturally, a review of the interpretations offered in this chapter.

### *3.5.2 Adaptive Planning in PAD? No, but exploration*

Instinctively, the notion of Adaptive Planning seems to be out of place in the PAD world, and even at second glance this assessment holds true.
It is obvious that a PAD task cannot be solved by invoking memories from our past and matching them to the current situation. In the ADE task used in the work of Hayes-Roth & Hayes-Roth (1979), this is to a certain degree possible. The verbal protocol produced by the participant in Hayes-Roth & Hayes-Roth' s (1979) study contains several statements that involve prior knowledge about the world and about the compatibility of errands, etc. E.g., he states that he wants do the groceries as late as possible, because otherwise the milk will go bad (pp. 278 - 279). The participant also assigns primary and secondary importance to the individual errands himself, according to his subjective views and presumably also based on his experiences. E.g. he decides that the errand to obtain medicine for the dog at the vet is "definitely a primary", although it does not say so in the instructions (p. 278).

The PAD world is a much more rigid place. Everything is pre-defined, from the times the appointments can be met to their priorities and the distances between them. Although he appointments themselves are realistic enough (who doesn't know the feeling of copying a book for one hour and a half?) the constraints of the task are rigid enough to evoke problem-solving behavior "from scratch". Participants have to find the schedule for a PAD task on their own, they can't simply retrieve it. Remember that for Altermann (1988) as well as for Schank & Abelson (1977) planning is problems solving; it is employed only if no script or previous plan for a situation is available. Viewed this way, the PAD world is a strenuous "worst case" for the participants, and so it should be - after all, one of its objectives is to measure planning and scheduling abilities, not memory capacity and swiftness of analogical reasoning.

Although Adaptive Planning can't be applied in the PAD world, it is worthwhile to discuss a close relative here: declarative learning. Declarative learning in the PAD world can occur in the form of accumulating knowledge about the feasibility of partial schedule. This notion is close to Logan' s (1988) theory of instance based learning, which states that problem solving is the result of the interpretation and exploitation of specific problem solving episodes.

While participants can't apply old experiences in order to solve a PAD task, they nevertheless gather new experiences during the solution of the task. In the course of their scheduling attempts, they find out which combinations of appointments are feasible and which are not. These experiences are certainly useful, because they help avoiding redundant states.
They also enable the participants to refrain from pre-calculating the possibility of moving to an appointment each time they have made a choice, because they already know for sure that certain partial schedules *do* work. This is a considerable relief for working memory, because the calculations involved in choosing an appointment in PAD can be quite straining, as we shall see in section x.

However, in the PAD world this kind of declarative learning comes at a cost. In order to learn about the feasibility of partial schedules, participants have to accumulate them. This means that some amount of schedule-modification is a prerequisite to declarative learning. These modifications can either happen because the participant has made a mistake (and is being told so by the system), or because the participant deliberately abandons partial schedules, *before*

In the first case, the knowledge that is accumulated is "negative" knowledge: Participants know how the solution *won't* look. This is useful knowledge, as the feedback of the system is always accurate. The schedules "learned" this way can be ruled out in future considerations and have not to be taken up again.
In the second case, the situation is more ambiguous. Should the participant take up schedules again, which he has already tried earlier, and abandoned?
On the one hand, the participant knows that a schedule works until the point at which he has abandoned it. That speaks in favor of trying it again. However, that is no guarantee that it will really work out. In fact, it was probably abandoned because it seemed to be unpromising.

Given the necessity to accumulate experience on the one hand, and the limitations of working memory[19] and time on the other hand, participants are faced two difficult decisions: "How often should I modify a schedule before a restart", and, later "should I take that schedule up again, or not". This of course prompts back to the introduction, where it was nonchalantly stated that, in the context of modifications, "moderation is the key". This statement is certainly true, and it certainly shows that there is no certain rule as to the amount of modification that is most supportive of optimal declarative learning.
A factor that determines the amount of modifications to a single schedule could be the presence of other appointments that look promising at the start of a schedule. If there are many, it is less risky to abandon a particular schedule; if there are few, more modifications of a schedule starting with a specific appointments can be expected.
A factor that determines the re-uptake of schedules starting with a specific appointment can be the reason why this schedule has been abandoned in the first place. If it was abandoned because the participant wanted to try something else, there is no reason to not try it again. However, if it was abandoned because the participant saw, looking ahead, that this schedule can't work (because it renders another appointment impossible), it is not reasonable to take it up again. Of course, the correctness of the look-ahead is crucial here.

That's why some participants may find it useful to use PAD as a helpful device, which enables them to test certain schedules (instead of simulating them mentally). Others may find this aversive.

---

[19] It is perhaps useful to point back to the number of "rational solutions" that was computed by Funke & Krüger (1993, p.6), which was 101 for PAD 4 and 388 for PAD 5. This gives a good impression of the scope of

This last point adds an interesting facet to the description of the "opportunistic pattern" made in the previous paragraph. The presence of many modifications, a high variety and many restarts may not only be viewed as an indicator of opportunistic processes. It can also be the result of a deliberate strategy of the participant: The strategy to (simply) explore the space of possible schedules directly by inputting them to the system, and to avoid extensive forward search or mental arithmetic.

This is an example of the ambiguity of patterns like this, and the resulting impossibility to definitely define the underlying process. For the remainder of the thesis I will therefore call this specific pattern the ***explorative pattern***, which includes deliberate as well as spontaneous exploration. A summary explanation of the explorative pattern is given in figure 3.1. An example Log file (imaginary) displaying an explorative pattern is shown in figure 3.2.

| ***Explorative Pattern:*** |
|---|
| Many modifications |
| Many Restarts |
| High Variety |
| Short Modifications |

*Figure 3.1: Summary of the explorative pattern*

(Appointments to be scheduled: Cafe, Secretary, Conference, Storehouse, and Post Office)

| | |
|---|---|
| • (Storehouse Secretary)<br>• (Storehouse Conference)<br>• (Conference)<br>• (Secretary Conference Storehouse)<br>• (Secretary Cafe)<br>• (Cafe)<br>• (Post Office Cafe) | *Number of scheduled appointments*: 5<br>***Variety:*** 5<br>***Restarts:*** 4<br>***Average modification length***: 2<br>***Number of modifications***: 7 |

*Figure 3.2: Example Log File showing an explorative pattern*

After this discussion of declarative learning in the domain of PAD, I will now describe a procedural view of skill acquisition. I will show where procedural learning can take place in PAD and, afterwards, show the connection of declarative and procedural learning processes in the PAD world.

## 3.6 ACT*: A procedural view of skill acquisition

In this section, I will discuss Anderson' s (1987) theory of skill acquisition. This theory is part of Anderson' s ACT* architecture (1983), a unified theory of cognitive performance, and as such must adhere to its constraints. However, as only the concept of skill acquisition is of immediate relevance to the present thesis, I will focus my discussion on that aspect.

Anderson' s theory of skill acquisition (1987) was devised with the objective to "account for differences in behavior by differences in experience" (p. 192). He claims that learning theories place an important and necessary constraint on models of cognitive skills, namely, that these accounts have to include plausible mechanisms that make this skill learnable at all. Anderson (1987) gives an extraordinarily concise overview of his theory in his abstract, which I will therefore partly quote:

> Cognitive Skills are encoded by a set of productions, which are organized according to a hierarchical goal structure. People solve problems in new domains by applying weak problem solving procedures to declarative knowledge they have about this domain. From these initial problem solutions, production rules are compiled that are specific to that domain and that use of the knowledge (p. 192).

An example of the application of a weak method to a new domain (paraphrased from Anderson, 1987, pp. 194 – 195) is the case of a novice subject, B. R., who learned to code function definitions in Lisp. In order to achieve this, she was provided with an introductory text on function definitions in Lisp, a specific example, and a template of such a function definition, which showed the general syntax, but left open spaces for the specific elements (see figure 3.1.).

| Template | Example |
|---|---|
| (**defun** <function name><br>(<parameter1><parameter2>...)<br><process description>) | (**defun t-to-c (temp)**<br>**(quotient (difference temp 32) 1.8))** |

Figure 3.3: The template and the example for coding Lisp-functions, as reported in Anderson (1987). The function "F-to-C" converts temperatures in Fahrenheit into centigrade.

B. R. used the weak method of analogy to solve that problem; i.e. she mapped her own function on the template, using the example.
This mapping involves multiple steps, between which the example function is inspected as a guideline. Accordingly, the first coding of a Lisp function takes some time[20].
However, Anderson (1987) reports an impressive speedup between the first and the second coding-trial[21], despite the fact that the second trial involved a more complex function (p.195). He explains this with a process called "rule compilation".
"Rule compilation" means the creation of new production rules that perform the steps that had to be established individually during the first trial in a single sequence. In the Lisp-context, that means that the example functions don't need to be inspected as often anymore. The weak method has changed into a task specific strategy.

This notion leads to an interesting prediction. Apart from the speedup between the first and second trial in learning experiments, it can also be predicted that there will be positive transfer between tasks that are structurally similar (i.e., in Anderson' s (1983) terminology, that have identical or similar goal structures), but no positive transfer between tasks that use the same declarative knowledge, but are structurally different[22]. The more similar two tasks are, the more transfer can be expected. Thus, cognitive skills are extremely task specific. He presents impressive empirical evidence that supports this prediction, from superficially different areas as text - editing, the development of geometric and mathematical proofs, and (once more) Lisp programming. A detailed account of this evidence is, unfortunately, beyond the scope of the present thesis. I will, however, briefly report an example from Lisp-programming.
Anderson (1987, p. 201) reports a study in which participants had to learn to evaluate Lisp-expressions, i. e. they were presented with the expression and had to predict to what value that expression would evaluate. As could be expected, participants got gradually better at doing this: they were faster in answering and they made fewer errors. In between these evaluation trials, participants were occasionally presented with the task to code Lisp-functions that would produce a specific output. This task uses the same declarative knowledge as the evaluation

[20] I can confirm this.

[21] Using data that were obtained with the CMU Lisp tutor.

[22] Anderson (1987) acknowledges that it is problematic to specify the productions (the "structure") that underlie two tasks, as "there is always the danger of fashioning production system models to fit the observed degree of transfer" (p. 198). He advises to consult different sources of independent evidence for specific productions, and

task, as Anderson (1987, p. 202) shows. However, performance in the coding exercise did not improve with time.

## 3.7 Transfer in the PAD world: An exploration of two specific PAD tasks

In the following paragraphs, I will first compare the necessary steps to solve two specific PAD tasks, PAD 4 and PAD 5, in order to analyze the possibility of Transfer between these two tasks. While transfer on the level of the appointments' criteria ("Macro level") is unlikely, on a lower ("micro") level, compilation of the mathematical steps involved in forward checking from Constraint Satisfaction Search can occur in PAD.

### *3.7.1 Criteria of the appointments*

We have already seen in the previous section that the method of adapting old plans to the current situation is not applicable in the PAD world. A similar thing is true for the analogy method mentioned in Anderson (1987). In the PAD world, participants are (usually) not presented with somebody else' s solution and then left with the option to try to solve their own task analogously.

But what about drawing analogies between two PAD tasks - in short, transfer? *If* I have found a good solution to the first of the two PAD tasks, can I use my knowledge about this solution to help me in the second task? I will address that question using PAD 4 and PAD 5, the tasks that were already introduced in chapter 2.

This may be a good moment to review the appointments for PAD 4 and PAD 5.

The figure shown below is identical to the figure in chapter 2. However, the solutions are added to the figure, as they will be referred to multiple times in the following discussion.

PAD 4

*Solution: (Administration-car-Storehouse-Conference-Secretary-Printing Office)*

- You have to be at the Storehouse between 10.00 a.m and 0.15 p.m. It will take you 10 minutes. *It's important.*
- Between 11.00 a.m. and 4.00 p.m you have to visit the Secretary. It will take you 10 minutes.
- You have to be at the Conference at 1.00 p.m, in the latest case. The Conference will last until 2.00 p.m. *It's very important.*
- You have to be at the Administration building at 2.30 p.m. It will last 90 minutes. *It's very*

- Between 10.00 and 4.00 p.m., you have to be at the Printing Office. This will take you 90 minutes. *It's very important.*

PAD 5

*(Solution: Printing Office-Conference-car-Office-Cafe-Central Office-Administration)*

- Between 1.30 p.m and 2.30 p.m., you can meet a customer at the cafeteria. The talk will last 30 minutes. *It's important.*
- Between 11.00 a.m. and 2.00 p.m you have to show up at the Office and deal with the files there. You will need 60 minutes for this. *It's very important*.
- You have to be at the Conference at 11.30 a.m, in the latest case. The Conference will last until 0.15 p.m. *It's important.*
- Between 10.00 a.m. and 4.15 p.m. you have to meet your boss at the Central Office. He wants to see you for 10 minutes. *It's very important.*
- Between 10.00 a.m. and 4.00 p.m., you have to be at the Administration. The work there will take 55 minutes. *It's important.*
- Between 10.00 a.m. and 3.00 p.m. you are to come to the Printing Office and copy a book. This will take 10 minutes

Figure 3.4. Appointments in PAD 4 and PAD 5, with solutions. There is only one complete solution in each PAD task.

One difference between the two tasks becomes obvious almost immediately: Although PAD 4 and PAD 5 "share" three appointments (the Conference, the Printing Office and the Administration), the times associated with the appointments paint a different picture for each of the two tasks. In PAD 5, the appointments all take place at similar times; in PAD 4, the times assigned to the appointments are rather different.

Consider, e. g., the appointments that start at 10 a.m. In PAD 4, these are the Storehouse, the Administration, and the Printing Office. The latest time to meet these appointments are 0.15 p.m. (Storehouse), 2.30 p.m. (Administration), and 4 p.m. (Printing Office).

In PAD 5, these appointments are the Central Office, the Administration, and the Printing Office. Their latest possible times are 4.15 p.m (Central Office), 4 p.m. (Administration) and 3 p.m. (Printing Office).

In PAD 4, the following linear sequence is suggested by the latest possible times, or urgency, of the appointments:

```
Storehouse, Conference, Administration, Printing Office,
Secretary.
```

On the one hand, this is nice, because it makes that particular task easier to solve, as the times restrict the choice of appointments for each slot in the schedule. For example, it is obvious that the Storehouse-appointment must be scheduled for an early time, and the appointments at the Printing Office and the Secretary can take place later.

However, this ordering can also be an obstacle in finding the solution. The crucial step in PAD 4 is to place the Administration *before* the Storehouse and take the car to the latter appointment. It can be difficult to see this, because it is not explicitly mentioned that the earliest possible time to meet the appointment at the Administration is 10 a.m. Instead, it says in the instruction: "You have to be at the Administration building at 2.30 p.m.". This is an ambiguous statement in English, and even more so in German. It can both be interpreted as "You must be there *precisely at* a specific time", and as "you have to be there at that time *in the latest case*".

The travel to the Storehouse by car is also risky; if the distance between Administration and Storehouse is reduced to a third, the participant is just in time. It is probable that participants calculate this in advance, but only sloppily, and thus arrive at the mistaken conclusion that a schedule that places the appointment at the Administration first wont work, as they will be too late for the Storehouse.

As already mentioned, there is more intersection between the appointments in PAD 5. This brings back to mind the remarks about constraints made earlier in this chapter. There it was said that a PAD task is most difficult for an intermediate "discriminating value" of the appointments' time-constraints. The most difficult case occurs if the appointments have identical time constraints but are, due to their duration and the distances between them, not mutually exclusive. In that case, participants can do nothing but search. PAD 5 is definitely closer to that "problematic" situation than PAD 4.

The crucial steps for finding the complete solution for PAD 5 are also different from PAD 4. The complete solution for PAD 5 is:

```
Printing Office, Conference, car, Office, Cafe, Central
```

The Printing Office is located far away from the starting point (the Office), and it is the sole appointment that is assigned no priority. Furthermore, it has "competition" from the Conference room, which is located next to it (Figure 2.1. or Appendix 1 can be consulted for a view of the PAD world). While in PAD 4, the information about the "urgency" of the appointment at the Storehouse has to be overcome in order to arrive at the correct solution, in PAD 5, the information that has to be "ignored" concerns the Printing Office' s low priority.

This points to some interesting similarities between the two PAD tasks. While they are perhaps not poignant enough to allow for a direct analogy, they nevertheless deserve highlighting here.

Firstly, both in PAD 4 and PAD 5, a relatively unlikely appointment has to be placed at the beginning of the schedule. Both the Administration (PAD 4) and the Printing Office (PAD 5) can be done until a relatively late time. Furthermore, in the case of the Administration, the earliest possible time isn't mentioned, and in the case of the Printing Office, the priority is low.

Secondly, in both cases exists an urgent alternative, which must be placed at the second position in the schedule, as it can't be met later. This is the Storehouse in PAD 4 and the Conference in PAD 5. However, as the Conference is a fixed appointment and can be visited at 11 a.m. in the earliest case, the situation is somewhat different.

Thirdly, in both PAD 4 and PAD 5, the remainder of the schedule can be found relatively easily after the difficult first choices have been made by linear chaining. We have already seen that in PAD 4. In PAD 5, when the Printing Office and the Conference are placed at the beginning of the schedule, the remaining appointments, ordered according to their urgency, are

```
Office, Cafe, Administration, Central Office.
```

Apart from the (basically trivial) permutation of the last two appointments (remember that such changes to the end of a schedule are supported by the PAD Interface), this sequence is exactly the solution.

The following two tables (table 3.1 and table 3.2) give an overview of the solutions for PAD 4

the current state of the schedule. (For example, in PAD 4, at solution step one, the Administration has to be chosen, which ranks highest according to the criteria start time and priority). The "competitors" (appointments that also rank highest according to that criterion) are placed in parentheses.

*Table 3.1. The solution to PAD 4 (leftmost column) and their ranking according to the given criteria. The appointments in parenthesis are appointments that also rank highest to the criterion at that point in the scheduling process. Trivially, the last appointment (Printing Office) has no competitors.*

| | Criteria | | | |
|---|---|---|---|---|
| | *Start time* | *urgency* | *priority* | *duration* |
| *Administration* | 1 (Storehouse, Printing Office) | 3 | 1 (Conference, Printing Office) | 3 |
| *Storehouse* | 1 (Printing Office) | 1 | 2 | 1 (Secretary) |
| *Conference* | 2 | 1 | 1 (Printing Office) | 2 |
| *Secretary* | 2 | 1/2 | 2 | 1 |
| *Printing Office* | 1 | 1 | 1 | 1 |

Table 3.2. can be read analogous to table 3.1. The solution to PAD 5 is set in the leftmost column.

| | Criteria | | | |
|---|---|---|---|---|
| | *Start time* | *urgency* | *priority* | *duration* |
| *Printing Office* | 1 (Central Office, Administration) | 4 | 3 | 1 |
| *Conference* | 2 | 1 | 2 (Cafe, Administration) | 3 |
| *Office* | 2 | 1 | 1 (Central Office) | 3 |
| *Cafe* | 2 | 1 | 2 | 2 |

| | | | | |
|---|---|---|---|---|
| | | | (Administration) | |
| *Administration* | 1 | 1 | 2 | 2 |
| | (Central Office) | | | |
| *Central Office* | 1 | 1 | 1 | 1 |

Some information in the tables is remindful of the critical remarks about the heuristic-analysis that is implemented in the PAD System (see chapter 2). In the same solution step, the same heuristic favors different appointments; e.g., in the first solution step the Storehouse, the Printing Office and the administration are ranked highest according to the criterion start time in PAD 4. The same appointment ranks highest according to different heuristics. This again shows the insufficiency of the overt criteria of the appointments as choice guidelines, and the difficulty to interpret the choice of an appointment as evidence for a heuristic that adheres to a specific criterion.

More importantly, tables 3.1 and 3.2 once more show that, after the "difficult" first choice of the first appointment (Administration or Printing Office), the criterion "urgency" allows one to find the solution. The appointment that has to be placed next in the schedule is always the most urgent one (without competitors). However, as we shall see in the next chapter, people do not seem to see that! They produce many modifications to a schedule, which makes it unlikely that they systematically attend to the urgency criterion (in that case, they would arrive at the solution much earlier).

In this context, I briefly (despite my fear of becoming redundant) want to comment once more on the work of Wolfram Schenck (2001), who has also discovered that implementing the choice of the next appointment in PAD as function of its overt characteristics is problematic.
In his thesis, Schenck (2001) discovered that his model of PAD, which he called EVA (for "Evaluation of Actions") had no difficulty fitting human performance in PAD 4, whereas the fir obtained for PAD 5 was mediocre (p. 91). In fact, a different parameter configuration had to be chosen for the two tasks, in order to fit them both sufficiently (pp. 86-87).
Schenck himself explains this "failure of EVA on Pad Task 5" (as he calls it somewhat exaggeratedly on p. 91) by referring to the necessity to place the appointment at the printing office at the beginning of a schedule in order to find the solution to PAD 5:

> Obviously, the appointment at the Printing Office lacks some of the characteristics that

> all scheduled appointments). Second, in regards to the earliest task starting time, it has two competitors with the same value (10 a.m.). Third, its urgency isn't especially high (latest task starting time is at just 3 p.m.). When one assumes that EVA relies on such overt characteristics of appointments for their evaluation, then the prospects of the Printing Office to be visited first are not very good. The other way round, the fact that EVA fails in PAD task 5, where a deeper assessment of the task configuration in obviously necessary, demonstrates that the evaluation carried out by EVA is most likely restricted to overt characteristics of the current PAD task and current state" (p. 91).

EVA indeed relies on such "overt characteristics" in evaluating appointments. This evaluation[23] is computed as follows:

> According to the enhanced evaluation, a maximum end score is calculated for every operator that could potentially be reached, if that operator were to be applied to the current situation, and the Delete Operator couldn't be applied afterwards. This constraint is important because without it, one would always be able to reach the optimal score (p. 43).

The "maximum end score" referred to in this quotation is, in turn, computed analogously to the evaluation in the PAD System (see also chapter 2): Each "very important" appointment that is carried out is worth eight points, each "important" appointment is worth three points, and every normal appointment one point.

It should now be clear why it was easier for EVA to find the solution for PAD 4, which requires to place a very important appointment before a "merely" important one (the administration before the storehouse).

On the other hand, while 56 % of the human participants placed the Printing Office at the first position of their final plans, the majority (39%) of the "subjects" simulated by EVA still prefer the important Conference (Schenck, 2001, p. 85).

Any PAD algorithm that assigns prime relevance to a single criterion will probably encounter the same problems EVA faces, and will have less difficulty the more the complete solution corresponds to a ranking of the scheduled appointment according to that criterion. IN PAD 4 and PAD 5, this criterion seems to be not "priority" but "urgency".

To sum up the Macro-section of this paragraph, I want to state that there are some surprising similarities between PAD 4 and PAD 5. However, it stands to reason of these similarities are sufficient to trigger analogical reasoning. PAD 4 and PAD 5, as these two tasks differ too much in terms of their appointments and the individual times allotted to them. The analysis of human performance in these two PAD tasks shall serve to qualify that judgement.

Let us now move on to the next ('micro"-) section of this paragraph, which is devoted to inspecting the rules that could possibly be compiled during a PAD session, and the methods from which they derive.

### *3.7.2 The Micro level: Constraint Satisfaction Search revisited*

I hold the view that both the methods that operate on PAD initially and the rules that are compiled afterwards are situated at a more basic, or microscopic, level than the rules that take the appointments' criteria into account. They are largely made of the mental arithmetic that is necessary to conduct simple forward search, which has nothing to do with the semantics of the appointments' criteria. In the following paragraph I will sketch such a "microscopic" set of rules that is independent of specific criteria of the appointments and discuss its implications for rule composition.

PAD can be stated as a Constraint Satisfaction Problem, as we have seen earlier in this chapter. There are some heuristics that can deal with Constraint Satisfaction Problems (Russell & Norvig, 1995, p. 104).

The ***most constraining variable heuristic*** finds the variable that is most constraining for the other variables and assigns a value to it first. In the PAD world, this heuristic would identify the fixed appointments, determine their time, and place the other appointments around them. Both in PAD 4 and PAD 5, this fixed appointment is the conference, which will always last until a fixed time (1.30 p.m. and 0.15 p.m., respectively). As this constraint cannot be changed, it constrains the placement of the other appointments, because the time-slot occupied by the fixed appointment must be kept free. The most constraining variable heuristic would sort the remaining appointments into the categories "before the fixed appointment", and "after the fixed appointment", and try to schedule them accordingly.

The ***most constrained variable heuristic*** would select the most urgent appointment and place it next in the schedule, because its time window is the most narrow, and it is unlikely that the appointment can be met later.

One could now argue that the two heuristics described above are also heuristics that order the appointments according to their criteria, and that there is no basic difference between them and the simple "ranking heuristics" that were criticized above. However, the crucial difference here is that the CSS heuristics do not establish an actual ordering but instead look at specific features of the *problem as a whole*. They identify a constraining appointment, no matter when it takes place and how important it is, and they select an appointment not because it is the most urgent one, but because it is more urgent than others are[24]. Moreover, these heuristics are merely sufficient establish a *preliminary* order among the appointments. the method of ***forward checking*** (Russell & Norvig, 1995, p. 84) is necessary to establish the choice of a next appointment –any choice, at any time.

Forward checking denotes the process of checking in advance whether the assignment of a value will fatally violate other variables' constraints. What does this mean in the PAD world?

It means that before I move to an appointment, it is tested whether that move will result in another appointment being "impossible"- that is, whether another appointment will now be impossible to meet.

This is the ultimate test that determines whether an appointment can be next in schedule or not. No matter how much a certain heuristic, from CSS or from somewhere else, may suggest the choice of some appointment –it must be made sure that this appointment will not stand in the way of another appointment, before that suggestion can be accepted.

My assumption that forward checking is the appropriate method for PAD (sensu "the initial weak method" in Anderson, 1987) is based on Anderson' s statement that "which weak methods can apply and how they apply are determined by what declarative knowledge in encoded about the problem domain" (p. 96).

In the PAD world, the initial knowledge encompasses the necessity to schedule appointments by moving to them, and the fact that these appointments are constrained and can't be met just any time. This is enough to include a "test" in the choice of the next appointment: the forward check. The mathematical knowledge of adding and subtracting time, translating hours into minutes etc., is a part of general knowledge. PAD requires no sophisticated knowledge about

the intrinsically nature of cafes or conferences to enhance performance. Indeed, what it probably requires most of all is a knack for mathematics.

The steps involved in the kind of forward checking sketched above are the following:

- Choose an appointment
- Determine the `current time`
- Add the distance between the current location and the chosen appointment to the current time to obtain the `current time + distance.`
- Add the duration of the chosen appointment to the `current time + distance` to obtain the `time after the potential appointment`
- Now, for every appointment that hasn't been done yet, do the following: Add the distance between the chosen appointment and the respective not-yet-done appointment to `the time after the potential appointment` and see if you arrive in time or not.

This "mental simulation" (cf. Dörner, 1989) sounds straining, and it is (see the Discussion in chapter 5). However, the steps sketched above "look ahead" only one step, but the performance of CSS algorithms improves with the extent of forward checking (Dechter & Frost, 2002). The option to use the car, which further enhances the complexity, isn't included in this sequence either.

The obvious complexity of a systematic forward search makes it implausible that Humans use it all the time. I will address that question later, but first, I will assume that humans do indeed apply forward search at least to some extent, and describe the consequences of rule compilation within this search method.

Firstly, the mathematical rules could combine to produce a speedup of the arithmetic. E.g., the computation of the "time after the potential appointment" could take less time, as the single steps are carried out in rapid series.

Secondly, the search for the remaining "not yet met" appointments could become more routine, e.g. when participants have figured out that houses in which these appointments take place are still white (the other houses are gray).

Finally, the distances and the times associated with the appointments remain the same throughout one PAD task, which means that retrieval (of addition and division facts) could enter the calculation.

In any case, the result would be quicker and better arithmetic. The rules would be specialists in pre-calculating times for the purpose of finding a next appointment in the PAD world, and as such context free and highly task specific.

The identical times that were needed by participants to solve various PAD tasks are no evidence against the formation of these rules. Anderson (1987, p. 199) reports the result of a study on skill acquisition and in text editing and explains that "(t)he actual time per keystroke in the execution phase did not decrease in the experiment, although there was some reduction in the number of keystrokes per edit, reflecting the acquisition of slightly more efficient procedures. This is exactly the pattern expected".

However, the fact remains that forward checking is a strain, and it is not plausible that humans apply it all the time. A relief can be the method of backtracking, which analyzes part decisions in the search in order to improve its progress. When we translate this statement into psychological terms, we arrive again at the importance of declarative learning and exploration. As Anderson (1987, p. 196) states: "The declarative knowledge encoded about (a) problem domain is again determined by the experiences of the learner: instruction, reading of text, examples studies, and so on".

In the final section of this chapter I will therefore explore how declarative learning and the acquisition of arithmetic skills is connected in the PAD world and must work together to produce good scheduling behavior.

## 3.8 Non modo, sed etiam: Procedural and declarative learning in PAD

In the PAD world, moderation indeed is the key. This statement, which should be the motto of this thesis, gains an additional meaning with regard to the connection between exploration and forward checking.

Consider exploration first. I have already argued that exploration is necessary to accumulate (preferably) negative knowledge about PAD, and the feasibility of partial schedules. This is important because it restricts the search space and is a relief from mental arithmetic. Indeed, PAD can be used just like a calculator, which receives partial schedules as an input and outputs the result: Either the schedule is possible –or not.

Unfortunately, however, there is a limit to working memory, and thus to the beneficial value

inputs them into PAD instead will soon have accumulated an impressive heap of partial schedules. Indeed, "pure" systematic exploration in a PAD task would equal an exhaustive "blind search" through the permutations of appointments –not a nice thought. But even with less systematic exploration, chances are that the participant will forget some of his experiences, or confuse them, drawing wrong conclusions. For example, if a participant wrongly remembers a schedule as feasible, he will not only end up with an incorrect solution attempt, but he will also gradually lose his trust in his own memory.

Forward checking, on the other hand, rules out certain solutions without the necessity to "go there and look". Although the method may be straining at first, it will improve with practice, and the skill of forward checking will enable the participant to recur to an exploration only if he has made a mistake in his calculation- in that case, the "exploration" will be a forced one.

Forward checking is also an efficient means to rule out certain schedules that start with a specific appointment a priori, because the very first appointment already renders another one impossible. In PAD 4 these appointments are the Printing Office and the Conference. In PAD 5, it is the Cafe. Interestingly, in the study described in the next chapter, I found some scattered partial schedules that start with the Printing Office or the Conference in the data for PAD 4, but none involving the Cafe in PAD 5.

However, not all appointments can be ruled out that easily. For some appointments, a hypothetical participant would have to look *a few steps* in the future to rule out certain schedules, not just one. This is not only straining, but also carries the additional risk of calculation errors. Here, previous exploration pays off tremendously. For example, if I have learned the impossibility of a partial schedule before, it makes no sense to attach that schedule to a new one and try that "combination". This specific use of memory is of course only possible when I have not accumulated too much of it- that is, if I can still overview it.

Thus, it can be seen that procedural and declarative learning can be, and are, most efficiently employed in connection with each other. Too great a preference for the one not only diminishes the quality of the other, but also the own – moderation is the key once more. A recommendation for trying to solve PAD could thus be: Try to check forward from the beginning, but when you haven't found a definite answer after two steps (in the future) – well, just go ahead and do it.

# 4 Modifications of Schedules

In study described in this chapter, the patterns of schedule modifications were analyzed, using empirical data of 43 participants that worked on PAD 4 and PAD 5. Specifically, it was tested if a high number of schedule modifications is due to the behavior that was described as "explorative" in the preceding chapter. To this end, the connection between the number of modifications, the number of restarts, the schedules' variety and the mean length of partial schedules was determined. While there was indeed evidence for the presence of this explorative pattern in both PAD tasks, it was moderated by the additional impact of the longest modification phase, showing that explorative schedules are also characterized by longer uninterrupted modifications of individual schedules.
The second objective of data analysis was to investigate how often participants modified their schedules without being prompted to do so by the PAD system (i.e. without being too late). If the percentage of such deliberate modifications increases between two PAD tasks, this is an indicator of an increase in forward checking, especially if the number of modifications itself remains equal.
Analysis of the data collected in this study showed that the number of modifications stayed the same in the two PAD tasks and that there was indeed a considerable increase in "deliberate modifications". This supports the assumptions about the acquisition of a PAD specific forward checking skill made in the previous chapter.
A third objective of the data analysis was to test whether some participants consistently exhibited an explorative planning style. This showed to be not the case: Participants who used many modifications in the first scheduling session didn't necessarily do so in the second. A tentative interpretation in terms of declarative learning is offered.

In the preceding chapter the explorative pattern of scheduling in the PPAD world was introduced. It was described as being characterized by many modifications, many restarts coupled with a high variety, and a shorter average length of the individual partial schedules.
Explorative behavior serves the purpose of accumulating experience and avoiding the dangers of calculation errors and the strain of forward checking.
Another assumption made in previous chapter concerned the method of forward checking as initial method to apply in the PAD world. The amount of forward checking bears an obvious connection to the number of "deliberate modifications". Deliberate modifications are modifications that occur without prompting by the PAD system, i.e., participants modify their schedules before they actually arrive at a dead end. Doing this certainly involves a certain amount of looking ahead.
However, the measure of deliberate modifications is ambiguous if it is viewed only in the context of a single task. As we have pointed out in chapter 3, "un-prompted" modifications can also signify spontaneous changes ("opportunistic" sensu Hayes-Roth & Hayes-Roth, 1979) that are not due to mental calculations. Therefore, it is more informative to analyze

of modifications, restarts and the variety (i.e.: the explorative pattern as a whole) remains equal. If an increase of deliberate modifications, relative to the total number of modifications, can be found under these conditions, this indeed indicates an increase in forward checking, and thus the acquisition of that specific skill. This analysis has been performed and will be reported in this chapter.

Before addressing the question whether the explorative pattern can really be found in empirical data, it is interesting to distinguish the explorative pattern from other patterns that could also yield a high number of modifications. These patterns are not mutually exclusive, but could be called subspecies of exploration.

## 4.1 Different Patterns

### *4.1.1 Many restarts/low variety*

One element of explorative scheduling as it is defined in this thesis is the coupling of many restarts and a high variety. As already explained, may restarts can also be the result of switching between a few appointments at the beginning of a schedule, which would yield a low variety. The differences between these two patterns can best be seen by using examples. They can be seen in figure 4.1.

| Explorative Pattern from chapter 3: ***many restarts, high variety*** | Pattern with ***many restarts, low variety*** |
|---|---|
| • (Storehouse Secretary)<br>• (Storehouse Conference)<br>• (Conference)<br>• (Secretary Conference Storehouse)<br>• (Secretary Cafe)<br>• (Cafe)<br>• (Post Office Cafe) | • (Storehouse Secretary)<br>• (Storehouse Conference)<br>• (Conference)<br>• (Storehouse Cafe)<br>• (Conference Storehouse Cafe)<br>• (Conference Storehouse Secretary)<br>• (Conference Post Office) |

*Figure 4.1: Explorative patterns with many restarts and a high/ low variety.*

The explorative pattern with the lower variety, as shown above, would suggest a more

completely fresh strategy, remindful of opportunistic planning (Hayes-Roth & Hayes-Roth, 1979).

### *4.1.2 Few restarts*

Many modifications could also be the result of many uninterrupted modifications to a schedule starting with a specific appointment. Again, let me contrast that pattern with the "original" explorative pattern described in chapter 3.

| Explorative Pattern from chapter 3: ***many restarts, high variety*** | Explorative Pattern with **few restarts** |
|---|---|
| • (Storehouse Secretary)<br>• (Storehouse Conference)<br>• (Conference)<br>• (Secretary Conference Storehouse)<br>• (Secretary Cafe)<br>• (Cafe)<br>• (Post Office Cafe) | • (Storehouse conference)<br>• (Storehouse Cafe Secretary)<br>• (Storehouse Cafe Conference Secretary)<br>• (Storehouse Cafe Post Office)<br>• (Storehouse Secretary Post Office)<br>• (Storehouse Post Office)<br>• (Storehouse Cafe) |

*Figure 4.2.Original explorative pattern, and many modifications/few restarts.*

The pattern of many modifications in connection with few restarts makes it obvious that many modifications can also be the result of the opposite of exploration: of being stuck. This is why the measure of restarts and variety are important to qualify the number of modifications in order to unambiguously identify behavior that can truly be called explorative. Another important measure is the length of the partial schedules. It has already been stated that short partial schedules are part of the explorative pattern. Let me give a brief explanation.

### *4.1.3 The importance of modification-length*

If the average length of the partial schedules is short, that means that somebody who was ready to modify them quickly created these schedules. (Remember that in the data format that was used for this analysis, a new "list" is created every time a modification of some kind is made). On the other hand, if the partial schedules are, on average, longer, this suggests more local changes to the end of the schedules –especially in connection with few restarts.

This last point truly highlights the importance of restarts, variety and schedule-length as *a pattern.* They can't be interpreted individually, but only in relation to each other.

One possible characteristic of the scheduling process hasn't been mentioned yet: The longest modification phase.

### *4.1.4 The longest modification-phase: Another measure of exploration.*

Up until now, it has been implied that the characteristics of many restarts and a high variety involve *quick* jumping between the partial schedules, i.e. that a single schedule is only modified a few times *in a row*.

This is, however, not compulsory. Many modifications in a row to a schedule can co-exist with many restarts and a high variety.

The kind of exploration characterized by many restarts, an high variety, and many "uninterrupted" modifications to the same schedule in a row would characterize a "super explorer", who not only inputs his "changes of mind" (restarts), but also many of the possible permutations to a single schedule he wants to check.

Another possibility would be a kind of reverse relationship between the variety and the longest modification phase: Either many modifications are due to (superficial?) exploration of many different schedules, or a more thorough exploration of a few select schedules.

For these reasons, the measure of the "longest modification phase" was included in the subsequent analyses.

## 4.2 Method

A study was carried out at the Department of Psychology, University of Heidelberg. A group of 45 participants, mostly students, took part in that study. Two of them were excluded from further analysis, as their data show anomalies that are probably due to a dysfunction of the PC's they worked on[25].

[25] Their data, as recorded in the log file, contained sequences like:"Conference, storehouse, Conference, Storehouse, Conference, Storehouse, (...)", without any deletions between them, which do not seem reasonable. As the participants seemed quite ordinary throughout the study, I assume this peculiarity is due to some minor

Participants were seated in front of a PC and solved first PAD 4 and then PAD 5. They received chocolates and (optional) course credit for their participation. They were instructed to read the PAD Instructions carefully. They were not instructed to find the only possible solution, but only, globally, to "try their best".

In the subsequent data analysis, the Log files created by PAD were transformed into the Lisp-like format introduced in chapter 2. The measures "number of modifications", "restarts", "variety", "average modification length" and "longest modification phase" were computed for each participant, both for the log file of his/ her performance in PAD 4 and in PAD 5.
The measures were operationalized as explained in chapter 2.
Additionally, the ratio of deliberate modifications to the total number of modifications was computed for each participant. It simply gives the percentage of modifications that occurred *before* the participants arrived too late at a location, relative to the total number of modifications created by that participant. Thus, if this ratio is high, that means that the participant was never too late, if it is low, it indicates more errors, and so on. This measure was computed by using a Lisp program that simulated the individual modifications.

To analyze the pattern underlying a high number of modifications, the correlations between the above mentioned measures were computed, both for PAD 4 and PAD 5. To test for the presence of an explorative planning style across the two tasks, the correlations between PAD 4 and PAD 5 were computed for the individual measures.

## 4.3 Results

### *4.3.1 Patterns*

The correlations between the measures under investigation in summarized in table 4.1 (PAD 4) and table 4.2 (PAD 5).

*Table 4.1: Correlations between the number of modifications, restarts, variety, mean of partial schedules and longest modification phase in PAD 4. The asterisks indicate significance at the level of 0,001 % according to Fisher's Z test for correlations.*

| | **modifications** | **restarts** | **variety** | **length** |
|---|---|---|---|---|
| Modifications | | | | |
| Restarts | .740*** | | | |
| Variety | .614*** | .794*** | | |
| Length | -.573*** | .-543*** | -.648*** | |
| Longest modification phase | .578*** | .-313 | .053 | -.343 |

*Table 4.2: This analogical table to table 4.1 depicts the correlations for PAD 5*

| | **modifications** | **restarts** | **variety** | **length** |
|---|---|---|---|---|
| Modifications | | | | |
| Restarts | .665*** | | | |
| Variety | .559*** | .836*** | | |
| Length | -.461*** | -.516*** | -.512*** | |
| Longest modification phase | .761*** | .098 | -.125 | -.248 |

The pattern of the correlation corresponds to the explorative pattern. Both in PAD 4 and PAD5, the number of modifications is highly positively correlated with the number of restarts, and variety, and highly negatively with the average length of the partial schedules. The measure "longest modification phase" is positively correlated with the number of modifications, in both PAD 4 and PAD 5.

The number of restarts is also highly positively correlated with the variety, and negatively with the average length of the schedules.

There is no inverse relationship between the variety and the longest modification phase.

The correlation between the number of modifications and the number of restarts/ the variety seems to weaken somewhat in PAD 5, although it retains its statistical significance. On the other hand, the correlation between the number of modifications and the longest modification

phase has increased. The same is true for the correlation between the restarts and the variety, suggesting an emerging either/ or relation among these measures (see discussion).

### *4.3.2 Deliberate modifications*

Table 4.3 shows the increase in deliberate modifications between PAD 4 and PAD 5. This increase is considerable. The high number of schedules containing no deliberate modifications, but instead only partial schedules that are impossible (mode of 0 % deliberate modifications ) in PAD 4 will be explained in the discussion.

*Table 4.3: The ratio of deliberate modifications to the total number of modifications, given in percentages, for PAD 4 and PAD 5.*

| | **PAD 4** | **PAD 5** |
|---|---|---|
| *Average (%)* | 41.7 | 59.6 |
| *Median (%)* | 44.0 | 58.0 |
| *Mode (%)* | 0.0 | 100.0 |

It is important to keep in mind that these statistics are only meaningful if the number of modifications (and, ideally, the explorative pattern as a whole) remains constant between the two PAD tasks. As the next paragraph shows, this is indeed the case.

### *4.3.3 Styles*

Table 4.4 shows average of the measures number of modifications, variety, mean schedule length, restarts, and longest elaboration phase for both PAD 4 and PAD 5, as well as the correlation between them. The range of each measure is given in the parentheses.

*Table 4.4 Correlations between the various measures from PAD 4 and PAD 5. Explanation is given in the text.*

| | *Number of modifications* | *Number of restarts* | *variety* | *Mean schedule length* | *Longest modification phase* |
|---|---|---|---|---|---|
| PAD 4 | 7.5 | 2.0 | 2.0 | 4.0 | 4. 0 |
| | (1 – 24) | (0-10) | (1 – 4) | (2 – 6) | (1 – 18) |
| PAD 5 | 7.9 | 2.3 | 2. 5 | 4.0 | 4.3 |
| | (1 – 24) | (0 – 10) | (1 – 5) | (3 – 7) | (1 – 9) |
| Correlation | .203 | .247 | .151 | .137 | -.152 |

## 4.4 Discussion

### *4.4.1 The absence of a scheduling style*

The analysis in this study has shown that the general pattern of the scheduling process seems to remain constant across the two PAD tasks that were used. A high number of modifications is always negatively correlated to the length of the partial schedules, and positively to the number of restarts and variety. This establishes the explorative pattern described earlier in this thesis. The magnitude of the correlations is astounding, if one considers the complexity of the analyzed data. The positive relation between the number of modifications and the longest modification phase also is stable across the two PAD tasks. This was no part of the original explorative pattern, but is no contradiction to it.

Considering the almost uncanny superficial resemblance between the data generated in PAD 4 and PAD 5 (table 4.3), the low correlations for the measures are surprising. Participants who needed few or many modifications in PAD 4 are not likely to need the same amount of modifying (and exploration) in PAD 5.

One plausible explanation for this finding is that it is possible to arrive at a fairly good solution in PAD without meeting all appointments (see chapter 2). As participants were not instructed to find a complete solution, they may have been satisfied with the first best schedule they found. Alternatively, they may have found the optimal solution right away, by sheer luck. In both cases, the positive "side effect" of a longer planning process, namely, the sharpening of the mathematical swiftness necessary for forward checking is lost to these participants, so some of them should face problems in the next task (PAD 5).

Participants who took longer (in terms of modifications) in PAD 4 have possibly gained at least some skills in the process that enabled them to perform better in PAD 5.

To test this assumption, the data created by the ten participants who needed the fewest modifications, and the ten participants who needed the most modifications, were submitted to further analysis. Tables 4.5 and 4.6 show the number of modifications needed by the two groups of participants in PAD 4 and PAD 5.

*Table 4.5: Modifications needed by the ten "best" participants in PAD 4 and PAD 5*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Modifications needed in PAD 4 | 1 | 1 | 1 | 1 | 1 | 1 | 2 | 2 | 3 | 3 |
| Modifications needed in PAD 5 | 1 | 4 | 10 | 24 | 2 | 5 | 4 | 5 | 4 | 1 |

*Table 4.6: Modifications needed by the ten "worst" participants in PAD 4 and PAD 5*

| | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Modifications needed in PAD 4 | 10 | 13 | 14 | 14 | 17 | 18 | 12 | 21 | 22 | 24 |
| Modifications needed in PAD 5 | 9 | 2 | 13 | 9 | 18 | 7 | 12 | 9 | 12 | 4 |

As can be seen, some of the participants who took few modifications in PAD 4 do indeed need substantially more in PAD 5. On the other hand, of the ten participants who needed many modifications in PAD 4, nine needed less modifications in PAD 5, the tenth needing only one modification more (18 instead of 17). As it has already been stated that a direct (semantic) analogy between the two PAD tasks is unlikely (see chapter 3), these improvements must be due to the enhancement of procedural skills. This shows that explorative behavior and procedural skill acquisition are not mutually exclusive.

Note, however, that the correlations in table 4.3 are not inverse. That means that the trend shown in table 4.4 is not strong enough to cause an inversion of performance. It is, however, strong enough to disrupt the pattern of "people who always take long", vs. "people who always get it right immediately.

### *4.4.2 Forward checking: the emergence of a skill*

The increase of deliberate modifications to partial schedules (and thus the avoidance of errors) is impressive. It is remindful of the dramatic statistics of one-trial learning reported by Anderson (1987, p. 195). This confirms the assumptions made in chapter 3 about the application of the method of forward checking in the PAD world, and its improvement.

Interestingly, the number of modifications, the number of restarts, and the other measures indicating explorative behavior remain stable across the two PAD tasks.

One plausible explanation for this finding could be that participants have acquired the basic skill of "looking ahead" in order to detect errors, but they only look ahead a limited number of

appointment into the PAD Interface. This is in fact reasonable behavior, as it enables them to "train" the skill of looking ahead, thus restricting the search space, while at the same time they make use of the benefits of exploration to relieve working memory (the current schedule is always displayed on the "Terminplaner").

### *4.4.3 Longest modification phase vs. variety: A subtle trend*

In tables 4.1 and 4.1, it shows that the correlation between the longest modification phase and the number of modifications increases between the two PAD tasks; it is higher in PAD 5. At the same time, the correlation between the number or restarts and the variety increase as well, while their correlation with the number of modifications decreases slightly. At the same time, the correlation between the number of restarts and the variety increases.

This could be interpreted as a subtle trend towards the following distinctions: A high number of modifications is either due to a longer modification phase of at least one schedule, or to the original explorative pattern.

While this may sound like a rather daring conjecture, it is nevertheless worth considering. A longer modification of individual schedules could indicate a more systematic approach to solving the task, which is still explorative, because it uses the PAD interface as a device to test schedules and gather experiences. However, given the order of magnitude of the correlations, the trend between PAD 4 and PAD 5 described in this paragraph is certainly too subtle to venture predictions for further development. PAD is too complex a task to expect clear-cut patterns after only two trials. Nevertheless, it points to the interesting option to expose participants to a longer series of PAD tasks, in order to monitor changes in the explorative behavior more closely.

Finally, let me counter-argue to the possible statement that, given the equal number of the modifications needed for PAD 4 and PAD 5, participants do not seem to have learned anything at all. This is, in fact, not true. Remember that PAD 4 and PAD 5 contain a different number of appointments to be scheduled (five appointments in PAD 4 and six in PAD 5). This makes PAD 5 more complex than PAD 4. Viewed in this light, the stability of the modification patterns across the two tasks can almost be interpreted as an improvement.

## 4.5 Summary

The number of modifications to a schedule in the PAD task is due to an explorative pattern,

with different appointments, short partial schedules, and a long series of modifications to at least one schedule. This pattern is stable across the two PAD tasks that were investigated in this study. However, the scheduling behavior of participants is not consistent across the two tasks; there is no planning style. This can be tentatively explained by a slow acquisition of procedural (forward checking) skills on the part of the explorers, nearly all of which (with one exception) need less modifications in the second PAD task, and by the fact that some participants who have arrived at a solution immediately in the first PAD task are performing considerably worse in the second one, due to their not having acquired task specific skills yet. The percentage of deliberate modifications to a schedule increases considerably between the two PAD tasks suggesting the relatively quick development of the skill of forward checking and error detection. This confirms the assumptions made in chapter 3 about the role of this method in the PAD world. However, the strategy of searching the through the possible schedules systematically (which would be signified by longer modification phases) seems to emerge much slower, if at all. Thus, the total number of modifications needed to find a solution remains stable. More expositions to PAD would be necessary to enhance the skill of systematic search in a way that would enable participants to truly exploit their newly acquired forward checking abilities.

# 5 Evaluation of Appointments

In the study described in this chapter, it was investigated how participants evaluate partial schedules (consisting of only one appointment) in PAD 4 and PAD 5, and, if they are presented with a specific appointment at the beginning of a schedule, which appointments they insert next. Participants were asked to give reasons for their evaluations and their choices. The evaluations of the appointments by participants in this study were compared to the choices made by participants in the study described in chapter 4. Despite some minor differences, the evaluations found in this study showed a strong correspondence to the choices found in the previous study. The analysis of the reasons participants gave for their evaluations and choices was performed in order to establish the impact of forward checking in the PAD task. While forward checking seems to play a role in the evaluation of individual appointments, the reasons given for the next choice suggest that different criteria are associated with the choice of different appointments. This suggests a rather unsystematic search for the next appointment during scheduling.

In the previous chapter, some empirical evidence was presented that suggested that participants working with PAD apply the method of forward checking, and that their application of this method improves between the two PAD tasks.

In the present study, a different approach to data analysis was chosen. Participants were asked to evaluate partial schedules, which consisted of single appointments from PAD 4 and PAD 5[26]. This evaluation was both numerical (participants were asked to give their score on a scale) and verbal (participants were asked to give a reason for their evaluation).
The evaluation of (other people's) schedules is, however, not the same as creating a schedule. Because of this, participants were also asked to pick a next schedule to insert into the partial schedules they are presented with, and give their reasons for that choice as well.

The main purpose of this study is to investigate more closely what criteria are crucial in participants' perception of the appointments in a PAD task. While the task analysis presented in chapter 3 concluded that content-independent forward checking is the most appropriate method to select the next appointment during a PAD session, it is still possible that participants attend to the more "obvious" criteria (e.g. urgency, priority) of the appointments after all.

This preference would not necessarily be a contradiction to the notion of forward checking (the respective findings presented in chapter 4 can hardly be refuted), but it could suggest a preliminary selection process that precedes it.

Thus, another objective of this study lies in investigating whether participants systematically adhere to one criterion in their evaluation and choice of appointments, or not. The explorative pattern found in chapter 4 suggests they don't, but, again, this study was conducted to qualify these assumptions.

Data analysis will proceed in the following steps.

First, the evaluation of the appointments will be reported and compared to the frequency with which these appointments were placed at the beginning of a schedule by the participants of the study described in chapter 4. Only if these correspondences are reasonably large, conclusion from this study can be drawn to the data patterns found in the chapter 4 study.

To permit a direct comparison, this analysis is restricted to the three appointments that are featured in PAD 4 and PAD 5. These appointments are the

`Printing Office, The Administration and the Conference.`

Interestingly, the "status" of these three appointments is quite different in PAD 4 and PAD 5: While the Administration is the first appointment in the solution of PAD 4, it has the undesirable quality of rendering another appointment impossible to meet if it is placed at the beginning of the schedule in PAD 5. The opposite is true for the appointments at the Printing Office and the Conference. While both of these appointments render another appointment impossible when placed at the beginning of a schedule in PAD 4, this is not true for PAD 5. And as we have seen, the appointment at the Printing Office is the first appointment in the solution to PAD 5.

Next, the reasons participants gave for their evaluations were analyzed by assigning their answers to categories. These categories were not pre-defined but extracted by abstracting from the participants written statements.

The same is true for the analysis of the reasons participants gave for their choice of the next appointments, which is analyzed afterwards.

For reasons of simplicity, this last part of the analysis (reasons given by participants to explain their next choice) was restricted to the PAD 4 group. However, the results of this explorative analysis are quite promising.

## 5.1 Method

### *5.1.1 Participants*

20 participants were assessed for this study, mostly students from the University of Heidelberg. They received sweets and (optional) course credit for their participation.

Participants were randomly assigned to two groups. These groups differ with regard to the PAD task participants had to work with and shall therefore be called "PAD 4 group" (N= 10) and "PAD 5 group" (N= 10). This decision was made to permit the identification of differences between the two tasks, or systematic correspondences between them with regard to the reasons given for the evaluation of the appointments. Apart from the context of the PAD task, the treatment was identical for the two groups. It is sketched below.

### *5.1.2 Treatment*

Participants were seated in front of an Apple Computer and presented with the "training" trial and either PAD 4 or PAD 5, to become more familiar with PAD. After they had worked with their respective task five minutes, they were told that they were now to evaluate the quality of the beginnings of schedules created by "other participants" (they were of course fictitious). They were told that these participants had worked with the same PAD task as they had themselves (that is, PAD 4 or PAD 5, respectively).

They were presented with a little booklet, containing five partial schedules, each consisting only of one single appointment at the start. In the PAD 4 group, these appointments were the ones that had to be scheduled in PAD 4, and in the PAD 5 group, they were the ones that had to be scheduled in PAD 5. The partial schedules were presented in form of a screenshot, which showed the "little square" of the other participant at the respective appointment, and the state of the PAD world.

This mode of presentation was chosen to enhance the comparability between this study and the one described in the previous chapter. This screenshot provides all relevant information for an evaluation (cf. Chapter 2): The other appointments that have yet to be done and their criteria, the current time, the distances. Participants in this study have exactly the same information as the participants in the chapter 4 study. Due to their introductory training session, they were able to interpret the screenshot correctly. The instruction sheets for the PAD 4 group can be inspected in Appendix 1.

Participants were offered a scale from 1 to 6 to score the partial plans, “6” being equivalent to the judgement “very promising”, “1” being equivalent to the statement “not promising at all”. In Addition, they were asked to state the reasons for their evaluations by simply writing them down.
After that they were asked which appointment they would insert next in the partial schedule they had just evaluated, and give their reasons for this as well.

The instruction sheets for the PAD 4 group and the PAD 5 group differed in one aspect. Participants of the PAD 5 group were simply presented with five partial schedules, each placing one of PAD 5’ s five scheduled appointments at the start. As it is not possible to place an appointment at the “Office” at the start of a schedule, this appointment was left out in this group.
Participants of the PAD 4 group were also presented with partial schedules consisting of one appointment, with one exception. Remember that the complete solution starts with the administration. In that case, the “difficult” (cf. Chapter 3) three steps of the solution were presented to the participants, i.e. the schedule they were supposed to evaluate was

`Administration, Car, Storehouse.`

The purpose of this minor manipulation was to see if participants of the PAD 4 group would evaluate the only solution of their task better than the participants of the PAD 5 group (who were just presented with its first appointment:

`Printing Office`.

If this should prove to be the case, this could be an indicator for the importance of taking the counter-intuitive first step, and for the easy detectability of the correct schedule after that step. However, if the partial schedule starting with the Printing Office (PAD 5) were evaluated equally good as the partial schedule starting with the Administration (PAD 4), the additional information given to the PAD 4 group would not be as crucial (and a neat argument made in chapter 3 would be refuted).

## 5.2 Results

As each of the 20 participants evaluated five partial schedules, the evaluations encompass 100 data points; that is 50 data points for the PAD 4 and PAD 5 group, respectively. Only the three appointments that featured in both tasks (Conference, printing office and administration)

were compared with respect to their evaluations, which means that 60 data points entered this comparison (30 per PAD group).

The selection of the next appointment is analyzed for each of the five partial schedules in both PAD groups and thus encompasses again the original 100 data points (50 per PAD group).

The reasons given for the selection of the next appointment is only analyzed for the PAD 4 group, and is thus based on 50 data points.

### *5.2.1 Evaluation of the partial schedules*

Table 5.1 shows how the appointments "Printing Office", "Administration" and "Conference" were evaluated by both groups of participants. Furthermore, it is also reported how many of the modifications created by the participants of the previous study[27] start with these appointments.

It can be seen that there is a correspondence between the scores and the number of modifications, which, however, is better for PAD 4 than for PAD 5.

*Table 5.1: Evaluations of three appointments in this study (average values), compared with the number of modifications starting with these appointments created in the previous study*

| | PAD 4 | | | PAD 5 | | |
|---|---|---|---|---|---|---|
| | Administration | Conference | Printing Office | Administration | Conference | Printing Office |
| Score (this study) | 5.6 | 1.5 | 2.3 | 2.2 | 3.9 | 4.9 |
| # modifications (previous study) | 126 | 2 | 12 | 60 | 71 | 72 |

### *5.2.2 Choice of the next appointments*

Table 5.2 reports the choice of the next appointment selected by the participants of the PAD 4 group. The appointments that were presented to them are listed in the leftmost column. The cells of the table contain the observed frequencies for the combinations of appointments.

*Table 5.2: Observed frequencies for the choice insertion of a next appointment into the partial plans that ere presented to the PAD 4 group*

| | Options for a next choice | | | | | |
|---|---|---|---|---|---|---|
| | Conference | Administration | Secretary | Storehouse | Printing Office | Delete |
| Administration | 10 | --- | 0 | 0 | 0 | 0 |
| Printing Office | 8 | 1 | 0 | 0 | --- | 1 |
| Conference | --- | 1 | 0 | 0 | 6 | 2 |
| Storehouse | 2 | 0 | 5 | --- | 3 | 0 |
| Secretary | 3 | 1 | --- | 5 | 0 | 1 |
| *Total* | 23 | 3 | 5 | 5 | 9 | 4 |

All participants of the PAD 4 group selected the "conference" as the next appointment after the partial schedule starting with the administration. One reason for this is the nature of this "partial schedule". Remember that it did not just contain the appointment at the administration, but the start of the complete solution, i.e. "Administration, Car, Storehouse". This is very probably the reason why the conference was picked so often as a next choice for the partial schedule starting with the administration, in the PAD 4 group –and not the storehouse

Table 5.4 shows the analogous information to table 5.2, this time for the PAD 5 group.

*Table 5.4: Observed frequencies of the next choices selected by participants of the PAD 5 group (N= 45, due to some missing values (participants didn't give a second choice).*

| | Options for a second choice | | | | | |
|---|---|---|---|---|---|---|
| | Administration | Conference | Printing Office | Central Office | Cafe | Office |
| Administration | 0 | 2 | 1 | 3 | 1 | 1 |
| Conference | 0 | --- | 6 | 0 | 1 | 2 |
| Printing Office | --- | 8 | --- | 1 | 0 | 1 |
| Central Office | 0 | 6 | 0 | --- | 0 | 2 |
| Cafe | 0 | 2 | 2 | 4 | --- | 1 |
| *Total* | 0 | 18 | 9 | 8 | 2 | 7 |

The frequencies of the delete option were not included in tables 5.4 and tables 5.5, because

schedule. In the PAD 4 group this happened slightly more often, which is why the operator was included in tabled 5.2 and 5.3.

### *5.2.2 Reasons given by participants for their evaluations*

The reasons that were given by participants in both PAD groups for their evaluations are shown in table 5.5.
In order to interpret this table properly, the transformation of the written statements that were produced by the participants into quantitative data must be explained.

In a first step, it was tried to abstract from the written statements in order to develop a system of categories. In this stage, it was tried to identify "classes" of reasons that participants mentioned. The aim was to find categories that were as mutually exclusive as possible. Another goal was to restrict the number of categories, in order to avoid a too complex analysis (that would be too arcane to be appreciated by the outside reader anyway).
I found that most statements made by the participants mentioned reasons that can be classified into the categories that are listed below. I am aware of the dangers that lie in defining such categories rather ad hoc and will address that issue in the discussion.

- *Other appointments*. This category encompasses statements about the possibility to meet other appointments in the future. The connection to forward checking is obvious. The other categories should be self-explaining.
- *Priority of the current appointment {that was to be evaluated}.*
- *Current Position*
- *Current time*
- *Priority of at least one other appointment*

Up until now, these reasons have little to do with evaluation. This element was added in the second step of the analysis of participants' statements. During the inspection of these statements, it became obvious that the same reasons were mentioned in connection to positive evaluations and negative evaluations. For example, "other appointments" can be the reason given for a positive evaluation, as in "I like that start to a schedule, because it will be possible to meet many other appointments". It can also serve to explain a negative evaluation, as in

“That start to as schedule is not good, because now many appointments can’t be met anymore”.
The same is true for the other reasons listed above. The priority of an appointment can lead to a positive evaluation: “This schedule is good, because I have already met an important appointment”. It can also lead to a negative one: “This schedule is bad, because this appointment is unimportant.” The other reasons can be transformed into negative and positive statements analogously.

The next goal was to determine how participants used the above-listed reasons to explain their evaluation of the three investigated appointments.
To do this, the following transformation was performed:

For each statement, it was coded whether a participant had mentioned the reasons listed above, and, if that was the case, if he had done so in a negative or a positive connotation.
Thus, each of the 5 categories listed above received a “score” from each statement. That score was

- “1”, if the reason was not mentioned
- “0”, if the reason was mentioned to explain why the partial schedule in question was bad (negative connotation)
- “2”, if the reason was mentioned to explain why the partial schedule in question was good (positive connotation)

Afterwards, these scores were summed up and divided by the number of participants.
The resulting scores for the categories, shown in table 5.5, can now be interpreted like this: A Score below 1 (close to 0) indicates that a reason from the respective category was mentioned to explain why the appointment in question was bad. A score above 1 indicates the opposite: A reason from the respective category was used to explain why the respective appointment looks good at the start of the schedule.
A score of (approximately) 1 can either indicate that the respective category plays no role in evaluating a particular appointment, or that participants’ judgements differ and the scores are averaged out.

*Table 5.5: Reasons given for the evaluations of the appointments for the PAD 4 and the PAD 5 group. The scores that were given to the appointments in the different groups are displayed in parentheses below them. Explanation on how to read the table is given in the text.*

| | PAD 4 group | | | PAD 5 group | | |
|---|---|---|---|---|---|---|
| | Administration (5.6) | Printing office (2.3) | Conference (1.5) | Administration (2.2) | Printing Office (4.9) | Conference (3.9) |
| Other appointments | 1.8 | .5 | .3 | .4 | 1.4 | 1.6 |
| Priority (Current) | 1.1 | 1.0 | 1.0 | 1.0 | .9 | 1.1 |
| Position | 1.0 | 1 | 1 | .9 | 1.5 | 1 |
| Priority (Other) | 1.2 | .8 | .6 | .6 | 1.5 | .9 |
| Current time | 1.5 | .8 | .6 | .6 | 1.5 | .9 |

This table shows that reasons that refer to "other appointments" mirror the evaluation of the appointments most closely. They are used to explain the "quality" of the administration appointment, and the lack of quality of the schedules starting with the conference and the printing office (PAD 4 group). The almost exactly reversed pattern is evident for the PAD 5 group.

The other categories do not offer such a clear correspondence to the actual score. The Categories "Current time" and "Priority (other)" are closest to such a correspondence.

This blurry picture is due to the fact most participants only give one reason for their evaluations and do not consistently give the same reason all the time. Moreover, it is simply due to the fact that different participants give different reasons for their evaluations. However, this topic belongs in the discussion section.

### *5.2.3 Reasons given by participants for their next choices*

The method that was used to extract the reasons participants gave for their next choices from their written statements was similar to the one that was used in the previous paragraph. In the first step, categories of reasons were established. These were found to be the following:

- Other appointments (again)

- Priority (of the chosen appointment; the subsequent criteria refer to the chosen appointment as well)
- Duration
- Position
- Urgency

Interestingly, these categories conform more closely to the criteria of the appointments that were given in the task instructions.

In contrast to the categories of evaluation presented in the preceding paragraph, the categories presented here are a priori "positive", i.e. *if* they are mentioned, they are mentioned in favor of the chosen appointment. For this reason, the coding of the categories was slightly different: A category was simply assigned a "1" if it was mentioned in favor of an appointment, and a "0", if they weren't mentioned. Afterwards, they were summed up and averaged.

Thus, if a category receives a score above 0 in the context of an appointment' s choice, it means that a reason from the respective category speaks in favor of that appointment.

Table 5.6 summarizes the results of this analysis.

*Table 5.6: Reasons given by participant of the PAD 4 group for their next choices. The leftmost column contains the "next" choices made by participants; the column headed "N" holds the number of times this choice was actually made by participants. More explanation can be found in the text.*

| | | Categories of reasons | | | | |
|---|---|---|---|---|---|---|
| | N | Other appointments | priority | duration | Position | Urgency |
| Conference | 24 | .3 | .4 | 0 | .15 | .25 |
| Administration | 3 | 0 | .3 | 0 | 0 | 0 |
| Secretary | 5 | .4 | 0 | .4 | .6 | 0 |
| Storehouse | 5 | .6 | .2 | 0 | .2 | .2 |
| Printing Office | 9 | .6 | .3 | 0 | .2 | .1 |
| Delete | 4 | .2 | 0 | 0 | 0 | 0 |

This table can be read in two ways. Firstly, the frequency with which reasons from the individual categories were mentioned at all can be seen in the columns. Inspecting the rows, one can identify the connection between the next choices and the categories.

For example, the category “other appointments” was mentioned most often to justify a choice and the category “duration” least often.

The choice of the “Secretary” as the next appointment seems to be connected to the criteria “position” and “duration” (remember that the secretary was only “chosen next” after the appointment at the storehouse; it is close to the storehouse, and, in PAD 4, lasts only 10 minutes). The choice of the “Storehouse” and the “Printing Office”, however, seems to be connected to the category of “other appointments”.

On average, one reason was given by participants to explain their next choices. This explains the subtlety of the effects.

## 5.3 Discussion

It cannot be denied that the results presented in this study can be called tentative at best. The design of the study carries some problematic aspects that I will discuss now. However, afterwards I will nevertheless comment on two interesting findings made in this study: The impact of “the other appointments” on both evaluation and choice, and the difference between the kinds of statements made by the participants in the evaluation and the choice context.

### *5.3.1 Problematic methodical aspects*

It is always a risk to define categories for verbal or written statements ad hoc. Chances are that another person, presented with the same data, would come up with a completely (or maybe not completely, but at least somewhat) different set of categories. I have offered no measure of the inter-rater reliability of the categories presented here, and I have the suspicion that, due to the eclectic nature of the coding (“1, 0, 2”), this reliability would not be overwhelming.

However, I do think that for the purposes of exploration, it is legitimate to propose a system like the categories described in this study. I do not hold the view that they implement a specific theory, nor do I claim that the findings presented in this study confirm or refute a particular theory. The purpose of the present analysis was simply to ask participants why do what they do, and for this purpose, the approach of “verbal common sense” is enough.

Nevertheless, there are interesting ways to modify the categories presented in this study and qualify them further. One way would (of course) be to involve multiple raters; another one could be to present participants with a list of reasons for their evaluations and let them check

Another problem that arose in this study is that participants were "forced" to give next choices for partial schedules they actually deemed "not promising". This biases the choice of next appointments, as well as the reasons given for them. Although participants had the opportunity to say that they wanted to delete the appointment they were presented with and start anew, they were not explicitly prompted to this possibility, and thus they may have bot used it each time they "felt like it". A possibility to cope with his situation would be to exclude appointments that were evaluated as "bad" from the analysis of the next choices.

### *5.3.2 The impact of the "other appointments": An indicator of forward checking?*

Both in the evaluation of appointments and the choice of the next appointments, the "other appointments" play a major role. What does this mean? It means that is it crucial for the evaluation of the appointments whether other appointments (that still have to be scheduled) can still be met. If that is the case, the appointment is evaluated positively, if it is not the case, it is evaluated negatively. This, to a certain degree, confirms the importance of forward checking in the PAD world, although this finding has to be qualified somewhat.

Firstly, the number of other appointments that are mentioned in the "evaluation part" of the study is, on average, just one. Table 5.7 holds the relevant information.

*Table 5.7: Number of other appointments mentioned by participants evaluating the appointments*

| | Appointments to be evaluated | | |
|---|---|---|---|
| | Administration | Conference | Printing Office |
| Number of other appointments mentioned (PAD 4 group) | .7 | 1.1 | 1.4 |
| Number of other appointments mentioned (PAD 5 group) | .9 | .3 | 1 |

Furthermore, the category "other appointments" also encompasses statements like "That schedule is bad, because I can't go to appointment x, *as it is already too late.*" This is not necessarily a sign of forward checking, as this would direct the judgement exclusively into the future, while the quotation simply states a present sate. It was sometimes hard to separate these two kinds of arguments, which is why I chose to compact them into one category. The dominance of this category is not contradictory to the presence of forward checking, but

say that the feasibility of other appointments is more crucial to the way a schedule is perceived that the criteria of the schedule that has already been produced. It would be interesting to see if this is also true for the evaluation of longer partial schedules.

### *5.3.3 Different reasons for evaluation and choice*

There are different categories of reasons given for the evaluation of partial schedules and the choice of appointments.

They are reviewed in table 5.8:

*Table 5.8:*

| Categories for the Evaluation | Categories for the choice of the next appointment |
|---|---|
| • Other appointments | • Other appointments |
| • Priority (current) | • Priority |
| • Priority (other) | • Duration |
| • Current time | • Position |
| • Current position | • Urgency |

While the reasons given for the choice of the next appointment focus on the criteria of that appointment, the reasons given for the evaluation suggest that participants assess the PAD world as a whole. The implications of the "other appointments" have already been discussed. The other categories for the evaluation point in the same direction, apart from the "Current Priority". The category "priority (other)" contains statements that explicitly mention another important appointment that can(not) be met, categories "current time" and "current position" contain statements concerning the fact that little/ too much time has be used yet by the partial schedule ("current time") and that the current position is close to/ too far away from other appointments.

On the other hand, this comprehensive assessment is lacking in the Categories for the choice of the next appointment. One possible explanation for this is that the evaluation of a schedule involves a different kind of reasoning that the justification of a choice. In the former situation, participants were asked to evaluate how *promising* they judge the appointments to be. It is obvious that this wording of the question evokes reasoning processes that probe into the future, checking what kind of situation the placement of a specific appointment at the beginning of a schedule could probably lead to. In the latter situation, participants have to justify something they have already done. Moreover, they are answering the last question on

the page. The criteria of the appointments (remember that they are visible on the screen, as they are displayed in the roofs of the houses) offer themselves as "quick and dirty" reasons to justify their choices. Another possibility is that, in different situations, specific appointments "stand out" and offer themselves as the next choice, because on single criterion speaks so much in their favor. An example for this would be the preference to insert the appointments at the Secretary' s office into the schedule behind the appointment at the storehouse, although it is not important, lasts only 10 minutes, and can be met until 4 p.m. (in the PAD 4 context). Another example would be the strong preference for the conference in PAD 5, the conference being a fixed appointment that has to be met rather early, thus standing out as well.

The preference for the adherence to different criteria in different situations is remindful of the pandemonium of Hayes-Roth & Hayes-Rothian planning specialists (1979): Pick the appointment whose criteria scream loudest: "Take me!".

As it has already been stated in the introductory passage to this chapter, this would not invalidate the notion of forward checking as the basic method in the PAD world (for more informal evidence on the application of that method, see paragraph below). Rather, it would suggest that participants restrain their search by applying some preliminary selection. Alternatively, participants may conduct systematic forward checking for a while, and if this doesn't yield a clear choice, they decide according to the criteria mentioned above. The only way to differentiate between these two explanations would be the use of verbal protocols.

This is not as far-fetched as it may sound now. Participants who gave me feedback about the experiment after the debriefing said that they found it straining to evaluate the partial schedules, since the only way to properly do that was to –forward check as many possible schedules starting with that appointment as possible. That left little energy for an elaborate justification of their next choices. That doesn't mean that participants didn't ponder them carefully; my observations during the experiment suggested that they did so, very much. However, the reasons given for the next appointments were often very general ("because it fits in the schedule", "because it is the only option"). This points to another reason for brevity, apart from exhaustion. Perhaps participants found the next appointment during the mental simulations involved in evaluating the partial schedules, and were so immersed in their thoughts that they found this choice too obvious to explain –which resulted in the "general" statements quoted above.

This surprised me, as I had expected participants to make spontaneous choices, without much

complexity of PAD (after having dissected it for this thesis for such a long time). The strain of the "non-interactive" paper-and-pencil evaluation of partial schedules prompted me once more to the beneficial characteristics of the PAD Interface, which enables participants to check partial schedules directly –to explore. A number of participants said that they had enjoyed the introductory PAD session (on the computer), during which they could try out some schedules, but had had a hard time during the evaluation part. One reason for this was that they found themselves unable to remember the beginnings of schedules they were currently simulating, and so had to start over again.

Many of the ideas about the importance of exploration were inspired by the post-debriefing discussions with participants in this study.

## 5.4 Summary

In this study, the criteria according to which participants evaluate partial schedules and choose the next appointment to be inserted into them were investigated. The feasibility of the remaining other appointments is an important factor in the evaluation of partial schedules, while more appointment-specific criteria play a role in choice. This latter finding can tentatively be explained as follows: either do participants use the criteria to restrict their forward checking, or they refrain to the use of the criteria when forward checking hasn't yielded a clear choice after some steps. These two explanations can only be differentiated by using verbal protocols.

# 6 Conclusions

Many interesting things were discovered in this thesis. Personally, I like the ideas about the connection between declarative and procedural learning, between exploration and skill acquisition the best, and sincerely hope that they will be used and scrutinized further, either by me or by another researcher.

As this thesis contains a lengthy introduction, and an abstract at the start of each chapter, I will not be as redundant to summarize my findings here. Instead I want to devote this last chapter to three implications of the ideas and findings presented in this work. The first is an attempt to answer the question "Is human scheduling any good?". The second implication concerns the interpretation of performance in the PAD task. The third implication concerns itself with possible cognitive models of the Plan-A-Day task. Let me discuss these topics in reverse order.

## 6.1 Implications for Cognitive Modeling

Some time (actually, some years ago even!), I tried to create an ACT-R model of human behavior in the Plan-A-Day task. Faced with the complexity of the empirical data, I was at a loss about how to detect some order in this chaos, how to find something simple and elegant that could be modeled. Eventually, I carried my ambition all the way to Pittsburgh, to the ACT-R summer school, where my mentor Jon Fincham and me finally produced a working model. This model[28] basically implemented various rules that selected an appointment according to different criteria (e.g. earliest start time, urgency etc.), along with a very basic forward checking method that looked one step into the future. The model also contained some simple backup strategies, for example, it took the car when it was too late, and a set of stopping rules that declared the scheduling process finished, even if no optimal solution was found (something participants also tend to do, and rightly so). Our model produced schedule processes that looked reminiscent of the human log-files in their variability and their chaos. Nevertheless, it contained no theoretically based elements, the choices it made were purely random. Its implementation was based on some common sense ideas I had about scheduling, and on the necessity to fit the human data at least superficially. It did not contain declarative or explorative learning either. Exasperated, I gave up.

[28] Available on the Homepage of the Department for Experimental and Theoretical Psychology, University of

During the writing of this thesis, I often thought back on that modeling effort, because many things I discovered and (finally) understood suggest quite interesting and elegant ways to model human behavior in PAD. The key discovery here is certainly the connection between declarative and procedural learning. Although the development of a Cognitive Model is not possible anymore, given the time frame of this thesis, I will nevertheless sketch at this place how such a model could be constructed.

A cognitive model should start out with two basic strategies (or production rules)[29], one that implements forward checking and one that implements "pure" exploration. The latter will just pick an appointment, place it in the schedule, and wait for feedback. If a negative feedback occurs, it will be stored in memory. The former will involve a series of productions that perform the arithmetic steps of forward checking. In the beginning, the exploration strategy will be preferred, as it requires less effort. However, as the negative feedback increases, the forward checking strategy will get its chance to be applied, leading to the proceduralization of the calculation-steps involved in forward checking.

The connection between exploration and forward checking could be implemented by the use of retrieval: The forward-checking productions could check whether there already was an experience matching the current calculation (e.g. a partial schedule that is not possible, the sum of some distance, the result of the "divide by three" option that is applied when the car is used...). If this is the case, the calculation can stop and the result ("this schedule will NOT work!") can be returned directly, leading to faster decisions and a lower number of errors. The fact that there always *is* exploration can be implemented by restricting the forward-check to one or two steps in the future. This guarantees the creation of new experiences, which is so important in the PAD world, while at the same time furthering the acquisition the crucial forward-checking skill. Thus, "pure" exploration gradually changes into forward checking that uses examples from memory to enhance its performance. A remaining unknown is the exact criterion for a restart. When and why so participants decide to try another schedule? This needs further investigation.

This is of course just a brief sketch of a model. The actual implementation would probably have to deal with more complexities. I nevertheless hope that this brief sketch has made the constructive value of the declarative/procedural connection made in this thesis somewhat clear.

---

[29] In this discussion I assume the Cognitive Model in question to be implemented in ACT-R, however, this is not

## 6.2 Interpretation of the Performance in PAD

This thesis offers an additional way to assess the performance of participants: The difference between the number of modifications needed for the first PAD task and the number of modifications needed for the second. This could qualify the more coarsely grained measures of performance. Remember that in PAD, it is possible to arrive at a good solution by accident as well as by "good scheduling". The number of modifications needed in the second tasks gives a hint whether the former or the latter is true. As we have seen, some participants arrived at their solution very fast in PAD 4, but needed much more modification in PAD 5. That pattern could indicate that the good performance in the first trial was due to a lucky accident. This allows for a direct transgression to the following and final question posed in this thesis:

## 6.3 Is human scheduling any good?

The answer to that question must be a most empathic "Yes". The quality of human scheduling is not s much a matter of the score people eventually achieve, but a matter of how they deal with the task, from beginning to end. It is a matter of the scheduling process.

I think the exploration employed by participants in the PAD world is a very intelligent and neat strategy. Participants exploit the option to test their schedules that is provided by the PAD Interface. That they nevertheless seem to employ forward checking in order to avoid dead-ends, and learn it so quickly between just two trials, is an additional finding that filled me with admiration. Remember that the PAD-Instructions mention five (PAD 4) and six (PAD 5) appointments, each associated with various criteria, each of which is somehow relevant in real life. Additionally, there is the option to take the car, which further complicates matters. Finally, a direct transfer between the two PAD tasks is not possible. Given this complex scenario, the participants' exploration, combined with their acquisition of the forward checking skill, shows almost heroic motivation and persistence.

# 7 References

Akyürek, A. (1992). *On a computational model of human planning.* In J. A. Michon & A. Akyürek (Eds.), Soar: A cognitive architecture in perspective. A tribute to Allen Newell (pp. 81-108). Dordrecht: Kluwer Academic Publishers.

Alterman, R. (1988). Adaptive Planning. *Cognitive Science (12),* 393 – 421.

Anderson, J. R. (1983). *The architecture of cognition.* Hillsdale, NJ: Erlbaum.

Anderson, J.R. (1987). Skill acquisition: Compilation of weak method problem solutions. *Psychological Review 94 (2),* 192 – 210.

Anderson, J.R. (1996). *Kognitive Psychologie*. Heidelberg, Berlin, Oxford: Spektrum Akademischer Verlag.

Anderson, J. R. & Lebiere, C. (1998). *The Atomic Components of Thought*. Mahwah (NJ), London (England): Lawrence Erlbaum Associates, Inc.

Carbonell, J. (1981). A computational model of analogical problem solving. *Proceedings of International Joint Conference of Artificial Intelligence*, Vancouver, Canada.

Carbonell, J. (1983). Derivational analogy and its role in problem solving. *Proceedings of the National Conference on Artificial Intelligenc*e, Washington, DC.

Dechter, D. & Frost, D. (2002). Backjump-based backtracking for Constraint Satisfaction Search. *Artificial Intelligence (136),* 147 – 188.

Dörner, D. (1989). *Die Logik des Mißlingens. Strategisches Denken in Komplexen Situationen*. Hamburg: Rowohlt.

Fikes, R.E. & Nilsson, N.J. (1971). STRIPS: A new approach of theorem proving to problem solving. *Artificial Intelligence, 2*, 189 – 208.

Fikes, R.E., Hart, P.E. & Nilsson (1972). Learning and executing generalized robot plans. *Artificial Intelligence, 3*, 215 – 288.

Funke, J. & Krüger, T. (1993). *"Plan-A-Day" {Computer-Programm}.* Bonn: Psychologisches Institut der Universität Bonn.

Funke, J. & Krüger, T. (1993). *"Plan-A-Day" (PAD): Ein Diagnostikum zur Erfassung von Planungskompetenz*. Manual zum Programm (unveröffentl. Manuskript). Bonn: Psychologisches Institut der Universität Bonn.

Funke, J. & Fritz, A. (1995a). *Über Planen, Problemlössen und Handeln*. In: Funke, J. & Fritz, A. (Hrsg.), Neue Konzepte und Instrumente zur Planungsdiagnostik (pp. 1 – 46). Bonn: Deutscher Psychologen Verlag.

Funke, J. & Fritz, A. (1995b). *Übersicht über vorliegende Verfahren zur Planungsdiagnostik*. In: Funke, J. & Fritz, A. (Hrsg.), Neue Konzepte und Instrumente zur Planungsdiagnostik (pp. 47 - 78). Bonn: Deutscher Psychologen Verlag.

Funke, J. & Krüger, T. (1995). *"Plan-A-Day": Konzeption eines modifizierbaren Instruments zur Führungskräfte-Auswahl sowie erste empirische Ergebnisse.* In: Funke, J. & Fritz, A. (Hrsg.), Neue Konzepte und Instrumente zur Planungsdiagnostik (pp. 97 – 120). Bonn: Deutscher Psychologen Verlag.

Funke, J. (2001). Dynamic systems as tools for analyzing human judgement. *Thinking and Reasoning 7 (1)*, 69 – 89.

Garey, M.R. & Johnson, D.S. (1979*). Computers and Intractability: A guide to the Theory of NP-Completeness*. New York: Freeman.

Ginsberg, M. L. (1993). Dynamic backtracking. *Journal of Artificial Intelligence Research*, 1, 25 – 46.

Hayes-Roth, B. & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science,*

Hertzberg, J. (1989): *Planen. Einführung in die Planerstellungsmethoden der Künstlichen Intelligenz*. Mannheim, Wien Zürich: BI Wissenschaftsverlag.

Hertzberg, J. (1995). *Planen aus Sicht der Künstlichen Intelligenz: Time for a Change*. In: Funke, J. & Fritz, A. (Hrsg.): Neue Konzepte und Instrumente zur Planungsdiagnostik (pp. 79-96). Bonn: Deutscher Psychologen Verlag.

Huchler, S. (1999). *Untersuchung eines Instruments zur Messung von Planungsfähigkeit im Rehabilitationsbereich.* (unveröffentl. Diplomarbeit). Bonn: Psychologisches Institut der Universität Bonn.

Kieras, D. E. & Meyer, D. E. (1997). An Overview of the EPIC Architecture for Cognition and performance with application to Human Computer Interaction. *Human-Computer Interaction, 12*, 391 – 438.

Klix, F. & Rautenstrauch-Goede, K. (1967). Struktur- und Komponentenanalyse von Problemlösungsprozessen. *Zeitschrift für Psychologie, 174* (3/4), 167 – 193.

Kohler, J. A., Poser, U. & Schönle, P.W. (1995). *Die Verwendung von "Plan-A-Day" für die neuropsychologische Diagnostik und Therapie*. In: Funke, J. & Fritz, A. (Hrsg.): Neue Konzepte und Instrumente zur Planungsdiagnostik (pp. 167 – 182). Bonn: Deutscher Psychologen Verlag.

Logan, G. D. (1988). Towards an instance theory of automatization. *Psychological Review, 22*, 1 – 35.

Newell, A. & Simon, H. A. (1972). *Human Problem Solving.* Englewood Cliffs (NJ): Prentice Hall.

Newell, A. (1973). *You can't play 20 questions with nature and win*. In: W. Chase (ed.), Visual Information Processing, New York: Academic Press.

Newell, A. (1990). *Unified theories of cognition.* Cambridge, MA: Cambridge University Press.

Pisanti, N. (1997) *A survey on DNA Computing*. Technical Report TR 97-07. Pisa: Department of Computer Science, University of Pisa.

Rattermann, M.J., Spector, L. Grafman, J., Levin, H. & Harward, H. (2001). Partial and total-order planning: evidence from normal and prefrontally damaged populations. *Cognitive Science, 25*, 941 – 975.

Russell, S. J. & Norvig, P. (1995). *Artificial Intelligence: A modern approach*. NJ: Prentice Hall.

Sanderson, P.M. (1989). The human planning and scheduling role in advanced manufacturing systems: An emerging human factors domain. *Human Factors, 31*, 635-666.

Shallice, T. (1982). Specific Impairments of Planning. *Philosophical transactions of the Royal Society of London (Biology), 298*, 199-208.

Schank, R.C. & Abelson, R.P. (1977). Scripts, plans, goals and understanding Hillsdale, NJ: Lawrence Earlbaum Associates Inc.

Simon, H.A. (1996). *Models of my Life*. New York: MIT Press.

Sipser, M. (1997). *Introduction to the theory of computation*. Boston: PWS.

Schenck, W. (2001*). A connectionist approach to Human Planning*. Heidelberg: University of Heidelberg {Diploma Thesis}

Wallach, D. (1998). *Komplexe Regelungsprozesse*. Wiesbaden: DUV.