Utilizing the Einstellung effect as a vehicle for evaluating training strategies in microworlds

Alexander Schreiber
Abstract

Utilizing the Einstellung effect as a vehicle for evaluating training strategies in microworlds

Alexander Schreiber

Based on their previous experience of solving problems, people can get stuck in a specific strategy to solve similar kinds of problems, which psychologists call the Einstellung effect. This thesis utilizes this effect in order to manipulate the usage of specific problem solving strategies in microworlds and, through the collected data, evaluates them against each other. An experiment was conducted to investigate these hypotheses, by training participants to use two different strategies to solve the same problem. The results indicate that the Einstellung effect, in combination with the microworld data, can be used to categorize strategies, but that the dynamic and complex characteristics of microworlds require more rigidly controlled environments to facilitate the usage of predetermined strategies. Controlling and recording interactions and progress through the microworld made it possible to compare the performance of different strategies and groups; thus, microworlds can provide a tool for capturing and analyzing data from more realistic lab studies.
Contents

List of Figures .................................................. V
List of Tables .................................................. VII

1. Introduction .................................................. 1

2. Purpose ....................................................... 3
   2.1. Delimitations ........................................... 3

3. Theory ....................................................... 4
   3.1. Operator selection ..................................... 4
   3.2. Problem solving ........................................ 5
       3.2.1. Heuristics and biases ......................... 5
       3.2.2. Hill climbing ................................... 6
       3.2.3. Mental models .................................. 8
   3.3. Einstellung effect ..................................... 9
       3.3.1. Impacts ........................................... 15
       3.3.2. Victims .......................................... 16
       3.3.3. Occurrence ...................................... 17
       3.3.4. Applications .................................... 18
   3.4. Microworld ............................................. 19
       3.4.1. Characteristics .................................. 20
       3.4.2. Microworld usage ............................... 21

4. Method ....................................................... 23
   4.1. Hypothesis ............................................. 23
   4.2. Design ................................................. 23
       4.2.1. Conditions ..................................... 25
   4.3. Microworld ............................................. 25
   4.4. Procedure ............................................. 26
   4.5. Pilot tests ............................................ 28
   4.6. Participants .......................................... 28
   4.7. Materials ............................................... 28

5. Results ..................................................... 30
   5.1. Scores .................................................. 30
   5.2. Critical times ......................................... 30
# Contents

5.3. Expert times ................................................. 32  
5.4. Instruction problems ....................................... 34  

6. Analysis ......................................................... 35  
6.1. Pilot data ..................................................... 35  
6.2. Training score ............................................... 36  
6.3. Room and time effects ..................................... 37  
6.4. Strategy analysis ............................................ 38  
6.4.1. Training ..................................................... 38  
6.4.2. Critical problem .......................................... 40  
6.5. Expert problem .............................................. 41  
6.6. Instruction problems ........................................ 42  

7. Conclusion ...................................................... 45  

8. Future work ..................................................... 46  

Bibliography .......................................................... 47  

Appendices  
A. Survey .......................................................... 51  
B. Participants ..................................................... 53  
C. Full results ...................................................... 55  

D. Microworld ......................................................... 57  
D.1. General instructions .......................................... 57  
D.2. Type I instructions ............................................ 58  
D.3. Type II instructions .......................................... 58  

# IV
List of Figures

3.1. Mathematical graph displaying the problem of hill climbing ............ 7
3.2. Relation of Einstellung problems to critical problems .................... 14
4.1. Chart of the experiment design ............................................... 24
4.2. Annotated screenshot of the microworld ................................... 25
4.3. Procedure of the experiment .................................................. 27
4.4. Map of both experiment rooms ................................................. 29
5.1. Plot of the distribution of the scores during the training phase .......... 30
5.2. Box plot that shows the scores both conditions received during the training phase ................................................................. 31
5.3. Plots of the distribution of the task time (a) and log(task time) (b) .... 31
5.4. Box plot that shows the task times for both conditions .................... 32
5.5. Plots of the distribution of the expert times (a) and log(expert times) (b) 33
5.6. Box plot that shows the expert times for both conditions ................... 33
6.1. Box plot that shows the task times of participants with a training score higher than one ................................................................. 36
6.2. Plots of all value changes in the sliders for all participants with a training score greater than one in the type I condition during the last three minutes of the training phase ......................................................... 39
6.3. Plots of all value changes in the sliders for all participants with a training score greater than one in the type II condition during the last three minutes of the training phase ......................................................... 39
6.4. Plots of all value changes in the sliders for all participants with a training score greater than one in the type I condition during the first five minutes of the critical problem ......................................................... 40
6.5. Plots of all value changes in the sliders for all participants with a training score greater than one in the type II condition during the first five minutes of the critical problem ......................................................... 41
6.6. Box plot that shows the expert score per minute ratio for both conditions 42
6.7. Plots that show the total deviation of the fill levels of all tanks from the type II condition during the training phase to the middle point of the tanks - the target area ......................................................... 43
A.1. Demographic survey ............................................................... 52
D.1. Annotated tank explanation ........................................ 57
D.2. Interactive example of how the interaction works ............. 57
D.3. Type I goal explanation ............................................. 58
List of Tables

4.1. Data collected during the experiment .............................. 24
4.2. Participants’ distribution over the conditions, times, and rooms .... 27
5.1. Results of a Welch’s t-test on the log(task times) ..................... 32
6.1. Differences in surrounding factors from the pilot tests to the experiment . 35
6.2. Overview of the task times in minutes from the pilot tests ............. 35
6.3. Results of a Welch’s t-test on the log(task times) of participants with a training score higher than one ......................... 37
6.4. ANCOVA with the task time as the dependent variable, the condition as a fixed factor, and the score as a covariate ......................... 37
6.5. ANOVA with the factors ‘start’ and ‘room’ on the logarithm of the task times ................................................................. 38
6.6. Mean and standard deviation of the expert score per minute ratio for both conditions ......................................................... 42
6.7. ANOVA with the factors ‘score’ and ‘condition’ on the expert score per minute ratio ......................................................... 42
B.1. Detailed information about the recruited participants .................. 54
C.1. Detailed information about the participants’ assignment to a condition and their performance ............................................. 56
1. Introduction

“Instead of the individual mastering the habit, the habit masters the individual.” (Luchins, 1942, p. 93)

Learning is at the heart of every new skill we develop – and we are learning new skills all the time. Thereby, learning is at the heart of everything we do and it starts with the very first time we encounter something new. Moreover, this first time is a unique situation that sets our path on how we will, at least initially, use these new skills and try to solve problems with them. This ability to perceive solutions and transfer knowledge onto similar tasks has made mankind great intuitive problem-solvers.

However, what happens if a problem appears to be similar to one that has been successfully solved before, but this time is either not at all or very inefficiently solvable by the previously applied strategy? In psychological experiments it has been shown that many people get stuck in their old way of thinking; therefore, they apply ill-fitting solutions because they incorrectly identify problems as being similar to previously solved ones.

We cannot change the fact that people always try to identify similarities between new problems and old experiences. They always strive towards making use of existing knowledge, as this creates less effort than trying to solve a problem from scratch.

Moreover, it is not always possible to clearly distinguish every unique problem from another, especially if we use generalized tools like the computer and its presentation layer, the screen. However, these changes are not necessary. Breaking down the ultimate goal into its core aspects and attributes, and then training people to solve these more basic and fundamental components will allow them to avoid getting stuck with ineffective strategies.

This approach is not new for domains outside of computer science. Professional musicians and athletes have known this for centuries. They spend thousands of hours training the core aspects of their skill. For the musician, these were, e.g., finger, foot, and breathing techniques, and for the athlete equally specialized and focused training on the main aspects of their sport. Combining these individual and isolated trained parts together to solve a larger and more abstract problem is then an almost intuitive step.

This effect also extends to more cognitive demanding skills that play a central role in education. Remember your math lessons in school? The teacher probably presented you with a problem and more often than not added a way to solve this problem. To make sure that you understand and remember this way of solving this problem, you repeated the procedure a couple of times; thus, you made the mental connection between this kind of problem and the strategy previously used to solve it. However, were you now to be presented with a problem that appears to be similar, but in fact is not, you...
would probably still have tried to solve it with your learned procedure. This effect of being blind to alternative and possibly more efficient strategies, when confronted with a similarly appearing problem, is called the Einstellung effect. Traditionally, Einstellung has been coined as a negative effect of learning transfer, but this thesis shall explore the advantageous applications that it can have.
2. Purpose

The purpose of this thesis is to investigate two hypotheses:

1. The Einstellung effect can be utilized to identify the usage of specific problem solving strategies.

2. Microworlds can be used to evaluate these strategies against each other.

2.1. Delimitations

The occurrence of Einstellung in motor processes and its effect in these areas will not be part of this thesis, apart from occasional usage for explanatory purposes.

Another part that is generally discussed when talking about Einstellung is rigidity and the ability to prevent Einstellung or to recover from it. As this thesis is exploring the option to explicitly create Einstellung, these topics will be only superficially covered.

Moreover, experts and expert knowledge will be used throughout the thesis to demonstrate the implications of Einstellung’s occurrence and how experts’ prior knowledge may make them inflexible, but the experimental part will be focused entirely on novice users. This means that there will not be any comparison between novice and expert performances and how they compare to each other with certain types of training.

The experimental part of this thesis will prime participants directly prior to their critical tests and long term priming effects will, therefore, not be studied.
3. Theory

The following sections will explore areas of related work, which provide the theoretical background required to study the influence of the Einstellung effect on strategies that people use to solve problems in microworlds.

3.1. Operator selection

Problems occur when people want to achieve a certain goal, but do not know how to solve it (Duncker, 1945). This comes costly for that person because cognitive strain is created when having to come up with a novel solution strategy. Due to a limited pool of solutions within a large sea of options, it is necessary for people to develop a strategy for selecting successful operators that work most of the time. An operator can be seen as a strategy or procedure that can be applied onto problems in order to possibly solve them.

One of the most predominant strategies that has been identified uses people’s experiences to select the operator that is most likely to be successful in the current situation (Luchins (1942), Reder (1987), Reder (1988), Thorndike (1932) (as cited in Lovett and Anderson (1996))). Thereby, generally speaking, a person confronted with a problem is likely to choose an operator that has successfully worked on past problems (Lovett & Anderson, 1996). This makes us analyze new problems to identify them as either being similar to previously encountered problems or being completely new. This use of analogy allows us to apply known operators to novel problems, which Gick and Holyoak (1980) therefore called a “hallmark of human intelligence”. They point out that this system of automatic recognition, however, is not trivial and it is not always obvious when two problems can be solved in the same manner. Correctly identifying an analogy is, nevertheless, a merit to successful problem solving because skill transfer often fails between similar domains and is virtually non-existent between different domains (Anderson, 2010). For this analogy selection to occur, Anderson (2010) states that the problem needs to be represented with salient cues that enable the solver to see which appropriate operator to apply. Making these cues salient can be done via different operator acquisition strategies like discovery and learning. Discovery leads to spontaneous associations between observations, actions, and results, whereby learning replaces the spontaneous associations with instructions, training, examples, and observations of others (Anderson, 2010).

Unfortunately, as the mind has limited capabilities and is always trying to reduce cognitive strain, it works similar to a “I’m feeling lucky” search on www.google.com, where the first search result will automatically be opened. This means that when an analogy is identified, the brain activates the operator sequence, which has previously been successfully used to solve that similar problem. Using this approach shows that humans do not
tend to think algorithmically, but heuristically – using the shortest possible way, meaning the way that creates the least cognitive strain, to a good enough solution (Anderson, 2010). Anderson (2010) presents one real life example of this, where someone who arrives via train at an unknown station and simply follows the masses in the hope that they lead her to the exit. He says that they, thereby, reduce the straining task of predicting values and probabilities by less straining operations that are easy to execute and appear promising. This is where the Einstellung effect comes into play, by providing another option to further reduce cognitive strain.

3.2. Problem solving

According to Newell and Simon (1972), problem solving can be described as trying to find a way to go from a present state to a goal state, where the solution is not trivial and the achievement of intermediate states may be required. If the solution would be trivial, the individual would not have to think about a way to solve the problem, which then cannot be described as problem solving (Duncker, 1945).

The focus of this thesis is to look at the specific influence of recent experiences – which are supposed to create an Einstellung effect on how people solve problems. The following sections will cover heuristics, biases, hill climbing, and mental models, to provide an understanding of how problem solving will work in this experimental study.

3.2.1. Heuristics and biases

Heuristics are mental shorts that are used to...

“[…] reduce the complex tasks of assessing probabilities and predicting values to simpler judgmental operations.” (Tversky & Kahneman, 1974, p. 38)

Kahneman (2013) later described heuristics as replacing a hard question with an easier one.

Heuristics are used in numerous different problem solving situations, but especially in those that are new and unfamiliar to a person. However, despite their wide and mostly automatic applicability, it is not guaranteed that the solution one arrives at will be correct (Colvin, Dunbar, & Grafman, 2001). Gigerenzer and Gaissmaier (2011) say that heuristics were often seen as intuitive reactions that were susceptible to errors and displayed irrational behavior. However, they also point out that heuristics incorporate many attributes that make them very suitable for everyday strategies. Besides being faster in reaching decisions than more complex and deliberate strategies are, they can also be more accurate when the solution is not required to be very specific and precise. Shah and Oppenheimer (2008) base the reduction in precision and decrease in execution time on the fact that heuristics rely on effort reduction by using all or a subset of the following approaches:

- Examining fewer cues
• Reducing the effort of retrieving cue values
• Simplifying the weighting of cues
• Integrating less information
• Examining fewer alternatives

These methods of reduction lead to people using less background information and instead focusing on information that is present at the current moment. Brehmer (2005) states that the participants who fail to use background information will most likely not uncover the essential characteristics of a complex and dynamic system (see section 3.4) that is presented to them. Moreover, those who fail at reaching their goal will, according to Brehmer (2005), display behaviors that prevent further learning from occurring. He gives two examples of how these behaviors could look like:

**Thematic vagabonding**  Jumping from one sub-goal to another

**Encystment**  Fixating on one sub-goal at the expense of other goals

Biases are the results of using heuristics and by understanding both of them, we could improve decision making in situations with limited knowledge (Tversky & Kahneman, 1974). The Einstellung effect is a bias that stems from the similarity heuristic, which judges how likely it is that an entity is part of one group or another, based on the similarity of the entity to those groups (Read, 2011). This entity can be a problem, which can then be judged to be part of certain groups of problems, based on their similarity to each other. The similarity heuristic can, however, as any other heuristic, produce wrong classifications. When incorrectly classified, the solution strategy that is connected to the group may not work for the new entity, which is not actually a member of that group of problems, although it may appear to be similar.

Another concept is functional fixedness. It describes a similar rigidity as Einstellung, but is connected to physical objects and the attributes that are traditionally associated with them (Duncker, 1945). One example is a lighter that is commonly used for setting something on fire. However, one can also use its hard form factor and the law of the lever to open bottles, which may not immediately occur to the camper who forgot to bring a bottle opener. This fixed relation between object and function can be seen as the physical analogy of the Einstellung effect.

**Einstellung effect**  Problem <-> Solution

**Functional fixedness**  Object <-> Function

### 3.2.2. Hill climbing

Hill climbing, difference-reduction method, repeat-state avoidance, and backup avoidance (Anderson, 2010) all describe the same heuristic that is often used to solve the
original Einstellung problems, which were presented as water jar tasks (Lovett & Anderson, 1996). It is often used in unfamiliar domains (Anderson, 2010); therefore, it is also used in experimental situations, where the participants have not previously encountered the material they are working with. Hill climbing describes the approach of looking at the current state and the goal state and subsequently choosing the next operator that will immediately take the participant closer to the goal state. By only looking one step ahead, the hill climbing heuristic is less demanding than other more advanced strategies, e.g., means-ends analysis. This could be the reason why Colvin et al. (2001) only observed the application of the hill climbing strategy, when their frontal lobe lesion patients attempted to solve water jar tasks.

However, it seems that this heuristic is used by a wide range of participants and is used to explain the initial approach of operator selection in various domains. Replications seem to indicate that although participants have no prior experience in a particular domain, they still do tend to be aware of actions that will bring them closer to their goal (Lovett & Anderson, 1996).

The problem with hill climbing is that, if it is strictly followed, one can get stuck in a sub-goal state without ever reaching the final goal state. This is demonstrated in Figure 3.1, where the current state is the coordinate (0,0) and the goal state is to reach the global maximum. The hill climbing strategy says that the next step is the one that takes us closer to the goal. This means that, outgoing from (0,0), we increase the x-value because that results in an increase in the y-value. However, arriving at the local maximum will lead the hill climbing strategy to fail because each further step would lead to a decrease in the y-value.

![Figure 3.1: Mathematical graph displaying the problem of hill climbing](image)

Transferring this to the water jar problem, one has to violate the hill climbing strategy in order to succeed in solving the problems (Colvin et al., 2001). The history of success (see subsection 3.3.3) plays here again a deciding role in whether a strategy is abandoned or strengthened. The more successful a person is with a strategy, the higher the chances that this strategy will be used for future problems.
3.2.3. Mental models

Mental models, as Norman (1983) describes them, are internally evolving models that humans create through interacting with a system. He describes several variations of models, but this thesis will focus on users’ mental models of the systems they are using. These models are theoretical concepts that users start creating when they first encounter a new system. Due to limited visibility and feedback, the models have certain drawbacks, which Norman (1983) summarizes in five points:

- Incomplete
- Limited simulation abilities
- Unstable
- Ambiguous
- Mental substitution

These attributes could arise partly due to the model’s connection to the individuals’ beliefs, values, and assumptions. They create them through interactions with the system and internal reflections on how these can be interpreted (Grösser & Schaffernicht, 2012). Mental models, therefore, provide a starting point to create hypotheses about the system that can explain how its underlying processes work. This information can then also be used to mentally simulate different manipulations of the system in order to predict what effect they would have on the current state (Ifenthaler, 2012). However, as Norman (1983) pointed out, people are generally not very good at simulating possible outcomes based on their mental models. That could be a reason why they tend to substitute physical actions for mental simulations, which Norman (1983) calls mental substitution.

Sterman (1994) investigated the creating and functioning of mental models of dynamic systems and discovered that it is hard for people to learn how these systems work. He connects this to several factors that make it difficult for people to form correct models. Some of these include the occasional inability to perceive the system’s feedback, time delays, and wrong assumptions concerning the system’s functioning. Inferring wrong mental models can be disadvantageous at best. If they occur within critical systems with little time to react, the consequences can be much more severe (Besnard & Greathhead, 2003).

While incorrect mental models generally seem to persist, Anzai and Yokoyama (1984) suggest that, with appropriate attention cues, these models can be changed to show a more correct representation of the encountered system. However, they argue that this depends on the degree of incorrectness and on how sensitive the participant is towards the cues. Too subtle cues with an unmindful participant and a too incorrect mental model will most likely not result in any change.
3.3. Einstellung effect

“[. . . when] the habit masters the individual – then mechanization is indeed a dangerous thing.” (Luchins, 1942, p. 93)

Einstellung effect is a bias in the choice of a plan (Luchins, 1942), which connects to the previously described operator selectors (see section 3.1). Historically, it has mainly been connected to Luchins (1942) and his experiments on students, whose task it was to use the representation of water jars in order to solve certain tasks. More information about the original experiments is provided in section 3.3.

The idea was to investigate the negative impact that preexisting knowledge can have on problem solving, which Luchins (1942) later referred to as a “mechanized state of mind”. He describes this condition as a “blind attitude toward problems”, which is induced when a person learns one way to solve a problem and later uses the exact same strategy on subsequent problems that can be solved by using a more direct and simple procedure.

This makes use of the previously discussed identification of analogy and the mind’s tendency to apply previously successful operators in order to reduce cognitive strain. Instead of saying that a general blindness occurs, Bilalić, McLeod, and Gobet (2008b) propose that alternative operators are suppressed when a familiar problem is identified that has a previously successful operator associated with it. This rigidity prevents people from thinking about, and thereby discovering, adaptive strategies and novel solutions that could either solve the problem or solve it more efficiently (Greenberg, Reiner, & Meiran, 2010).

The underlying principles of Einstellung are a step towards automated actions, which make expert performances more efficient, but also open up the possibilities for misguided actions that lead to undesired results (Bilalić et al., 2008b). Due to their automated nature, these associations are involuntarily made and require explicit attention redirection, which creates cognitive strain, to take other strategies into consideration. As John Maynard Keynes put it,

““The difficulty lies, not in the new ideas, but in escaping from the old ones, which ramify, for those brought up as most of us have been, into every corner of our minds.” (Keynes, 2006, p. 5)

Humans do not work algorithmically and when the blinding effect of Einstellung comes into play, people consider the critical aspects of problems less and less. They are led by the mechanical application of the activated operator that came from their previous experience. This tendency to view the problem only in the perspective of the activated operator hinders the mental ability to search for more efficient and effective strategies (Luchins, 1942).

For Einstellung to occur, it is necessary that a problem is incorrectly categorized as being similar to a previously encountered problem. Furthermore, the problem needs to require different solution paths, which can make the automatically activated solution strategy ineffective or inefficient. It is especially the negative connotation of Einstellung, which gave it its distinct name. Generally, one can say that retardation or facilitation
Einstellung effect occurs depending on if the subsequent problems are either appropriately or inappropriately solved with the trained strategies – but only the inappropriate case is considered to be Einstellung.

**Naming issues and similar effects**

The Einstellung effect has received its name over 70 years ago from Luchins (1942), who chose the German word for ‘attitude’ or ‘set’ to describe this rigid blindness when people attempt to solve a problem. There are, however, numerous different names and terms that researchers use to either describe the same phenomenon or phenomena that have more overlying effects. Two lists are provided to get a broad overview of what can be considered to be Einstellung and what can also describe a number of other effects and biases. The lists may not be complete and some may argue for a different separation, but they have helped me to find my way to related research articles.

Terms that describe the same phenomenon

- Einstellung/ Einstellung effect (Luchins, 1942)
- Set/Set effect (Woltz, Gardner, & Bell, 2000)
- Habituation (Luchins, 1942)
- Problem solving set (Ellis & Reingold, 2014)
- Strong-but-wrong error (Reason, 1990)

Terms that describe overlying or similar phenomena

- Functional fixedness/fixedness (Duncker, 1945; VanLehn, 1988)
- Negative transfer (Woltz et al., 2000)
- Capture error (Norman, 1981)
- Cognitive rigidity (Greenberg et al., 2010)
- Confirmation bias (Bilalić et al., 2008b)
- Representativeness heuristic (Tversky & Kahneman, 1974)

Einstellung is often considered to be part of two problem solving sets or set effects, where Einstellung is one of them and functional fixedness or fixedness the other (Greeno, Magone, & Chaiklin, 1979; VanLehn, 1988).
Negative connotation

It is definitely unusual to speak about the advantageous application Einstellung because the effect explicitly explains a negative learning effect. Researchers called it ‘blinding’ (Greenberg et al., 2010), a ‘disadvantageous set’ (Bilalić, McLeod, & Gobet, 2010), an ‘inappropriate schema’ (Sweller & Cooper, 1985), and said that it creates a ‘negative impact’ (Ellis & Reingold, 2014), ‘negative transfer’ (Woltz et al., 2000; Ellis & Reingold, 2014), ‘negative aspect’ (Sweller, 1989), and a variety of other negatively related concepts. Throughout its history of studies and experiments, everyone was keen on pointing out and researching its negative effects on transfer in skill performance and problem solving – starting with Luchins (1942)’ water jar experiments in 1942.

This report does, however, makes use of the value of Einstellung as a vehicle to identify the use of certain strategies across problems. This is especially useful in experimental environments, as its occurrence suggests that learning transfers did, in fact, occur. The critical question that then remains is whether these transfers were positive or negative and, thus, depend on other variables like tasks, environments, and instructions.

Advantages

Having a bias like the Einstellung effect has, nevertheless, also its positive sides. As previously mentioned, having to find a novel solution strategy to a problem creates cognitive strain. Being able to invest this effort and, thereby, solving the problem is something that separates humans from animals. However, if humans would ponder every question and problem that occurs – no matter how often this problem occurs and how similar it is to similar problems – they would waste a lot of precious energy. This ability to recognize similarities and analogically and automatically apply strategies reduces the cognitive strain we need to invest in reoccurring everyday problems. This freed up time and the available resources can then be invested to deal with more complicated situations (Luchins, 1942; Bilalić et al., 2010). Moreover, even when a novel solution is required, it is seldom the case that this solution needs to be completely perfect. Instead, a good enough solution may often suffice, as the best solution can be very hard or even impossible to reach, while only providing minimal benefits (Bilalić et al., 2008b).

This means that if a problem is correctly classified and a previously learned routine exists, the response is immediately available, quickly applied, precise, and increases our certainty in the situation (Luchins, 1942; Sweller, 1989; Schultz & Searleman, 1998; Bilalić et al., 2010). Being able to learn from previously encountered problems and applying the solution strategy analogically to other problems is a major part in skilled problem solving performance and in increasing this skill (Sweller, 1989; Woltz et al., 2000). Moreover, the likelihood of these positive effects occurring is increasing with increased similarity between the current and previously encountered problems (Woltz et al., 2000). Ironically, training with a single strategy – the basic principle that the Einstellung effect builds upon – is, according to Crooks and McNeil (2009), said to free up resources to allow for the following problem solving processes to occur:

- Noticing novel problem features
• Generating new solution strategies
• Extracting meaning from unfamiliar problems
• Suppressing inappropriate solution strategies that have been retrieved many times in the past

Especially the last item points directly to the Einstellung effect, which is created by reoccurring training situations; although it has been shown that people do not learn from the inappropriate use of strategies (Sweller, Mawer, & Ward, 1983; Bilalić et al., 2008b; Ellis & Reingold, 2014). The experiment in this thesis will explore both, the positive and negative results that the occurrence of Einstellung can create.

Origin
To understand the Einstellung effect, its implications and effects on experimental studies, it is beneficial to look at its origin – Luchins (1942)’ water jar tasks. Here is Luchins (1942)’ description of the original task:

“The experimenter told the class that its task was to figure out on paper how to obtain a required volume of water, given certain empty jars for measures. To illustrate this principle we presented Problem One. The subjects were asked for the solution, and the method of solving the problem [a 29-quart jug three times filling a 3-quart jar], was then written on the blackboard.” After this, Problem Two was put on the blackboard, [a 21-quart jar, a 127-quart jar and a 3-quart jar] get 100! After 2 1/2 minutes the subjects were asked for their solutions. The answer was then illustrated in both a written and verbal form; viz. [the 127-quart jar one time filling the 21-quart jar and then three times the 3-quart jar] and verbally ’One fills the 127-quart jar and from it fills the 21-quart jar once and the 3-quart jar twice. In the 127-quart jar there then remain the 100 quarts of water.’ Without any further interruptions the other problems, in succession, were presented one at a time [...]” (Luchins, 1942, p. 2)

The idea was that the successive repetition and usage of the same solving strategy would lead to people using it no matter if it was less efficient than other alternatives or not working at all. ‘Critical tasks’ or ‘critical problems’ were those tasks where other, more efficient strategies existed, but where the trained strategy was also possible. ‘Extinction tasks’ or ‘extinction problems’ were tasks, where only an alternative strategy, different from the training strategy was possible.

In general, the results showed that the participants were very likely to continue using the trained strategy, even when other more efficient strategies were available. Moreover, the test participants were unlikely to discontinue their usage of the inferior strategy, even when confronted with a problem that required another strategy to solve it. Furthermore, a majority of these participants was completely unable to solve the extinction problems. In comparison, test participants who were not trained towards a certain strategy solved
the great majority of extinction tasks and almost always used the more efficient strategy when confronted with critical tasks.

One of the water jar experiment’s main results include the observability when obvious solution strategies are missed, due to repetitive training sequences that primed a participant towards another strategy (Greenberg et al., 2010). This, combined with the fact that the water jar tasks require a unique problem solving strategy (Colvin et al., 2001), makes it a useful tool for experimental tests that are supposed to involve more cognitive strain and an adaptive problem solving attitude. The specific approach of the experiment can be described as requiring a “manipulation of defined material […] within a state space […] according to a set of predetermined rules” (Colvin et al., 2001, p. 1130). Generally speaking, going from a starting state to a pre-defined goal state. To be able to generalize the results of any experiment based on water jar tasks, it is, nevertheless, necessary that no solving patterns exist prior to the test. This ensures that the presented strategies are genuinely new and show an original approach that is only influenced by other independent variables (Wiley, 1998).

**Tasks**

When conducting Einstellung experiments, there are three different types of problems, where the participants get the task to solve them (Luchins, 1942):

- Training, Set, or Einstellung problems
- Extinction problems
- Critical problems

The main purpose of extinction problems is to confront the participant with a problem that cannot be solved by the strategy learned through the Einstellung problems. Subsequent critical problems could then be used to reveal if the participant falls back into her old patterns or if a recovery from the Einstellung effect can be observed.

The Einstellung problems are the initial starting point for the participants and present them with a problem that is to be solved. They may only have one solution because the strategy of solving the problem is what the participants are primed on. In combination with the critical problems, where the problem can be solved using either the strategy from the Einstellung problems or a more efficient strategy that was not previously encountered, one can observe how fixated the participants have become on one strategy. This is why it is necessary for participants to perceive the Einstellung and critical problems as being closely related (see Figure 3.2). If they incorrectly identify the two problems as being related, they will try to apply the same solution strategy; thus, they show the Einstellung effect.

Apart from the perceived relatedness of Einstellung to critical problems, the amount of Einstellung problems that are necessary for an Einstellung effect to occur has rarely been discussed. Crooks and McNeil (2009) argue that this can be due to Luchins (1942)’ original research, where he reported that additional training problems had no effect
on the Einstellung effect. They replicated Luchins (1942)'s experiment with six different conditions, where each group received a different amount of training. They observed that the participant's flexibility decreased linearly with the number of additional Einstellung problems, but that an effect is already present after a single problem.

Additionally, Luchins (1942) and Bilalić et al. (2008b) looked at the effect of time limits and concluded that more time does not prevent the Einstellung effect from occurring. On the other hand, reducing the available time and, thereby, inducing a stress factor strengthens the effect (Luchins, 1942; Schultz & Searleman, 1998).

The transition between the problems and, thus, the elapsed time before the participant encounters a critical problem is also something that Luchins (1942) discussed in his original water jar experiments:

“It is of the character of an Einstellung experiment in which a habit does not act simply because of some previous experience at some previous time, but a perseverative tendency is actually established in the very experimental situation, [...]” (Luchins, 1942, p. 28)

Regarding this, Wiley (1998) provides additional insights by arguing that the Einstellung effect among novice users may be breakable by waiting for it to decay over time. Combining this with Luchins (1942)'s statement suggests that if one wants an Einstellung effect to occur, the time between task problems should be kept as short as possible.

**Replications / Adaptations / Other domains**

The Einstellung effect's impact would be far smaller, if it would only be observable for experiments similar to the water jar tasks conducted by Luchins (1942). There are, however, many replications of the water jar tasks that have been conducted over the past decades, which makes the effect a very robust finding. These include more direct observations of the inflexibility of thought after training sequences (e.g., Atwood and Polson (1976), Chen and Mo (2004), Delaney, Ericsson, and Knowles (2004), Lippman (1994), Lovett and Anderson (1996), McKelvie (1990), Woltz et al. (2000) (as cited in Bilalić, McLeod, and Gobet (2008a))), whereas some especially noteworthy examples
Impacts Theory

include the works of Bilalić et al. (2008a), Vallee-Tourangeau et al. (2011) and Colvin et al. (2001).

Bilalić et al. (2008a) have conducted a series of extensive experiments, where chess players of various skills were confronted with more or less obvious solutions to a current problem. The results showed a correlation, where more experienced players were less affected by the Einstellung effect than novice players. Vallee-Tourangeau et al. (2011) introduced more interactive elements to the experiment in order to see if tangible environments changed the observability of Einstellung and indeed, the effect size was lower where participants were able to see and manipulate actual water jars. Finally, Colvin et al. (2001) also used the Einstellung effect, as they argued that the “unique combination of problem solving and planning strategies [allows for] a more precise identification of frontal lobe lesion patients’ cognitive deficits” (Colvin et al., 2001, p. 1129).

These are only a few examples, but the range of domains, where the Einstellung effect can occur, are almost endless because it is directly coupled to learning transfer and how previous training or priming occurrences can influence this transfer.

If one, however, wishes to build on replicated studies of the Einstellung effect, it is necessary to have a clear understanding of what part of Einstellung was tested. Looking at Bilalić et al. (2008a) and the study of Einstellung effect in chess players, it was a limited focus in comparison to the original water jar tasks. The water jar tasks were designed to produce the effect of Einstellung, whereas Bilalić et al. (2008a) only described Einstellung as first seeing an obvious (or well known) solution and then not continuing to look for other, less well-known solutions. In the chess experiments, the optimal solution was already all the time existent, but in the water jar tasks the easier (or more optimal) solution was added in the critical and extinction problems. It was not that the participants did not find the easier solution, but that they had to find the much harder (or less well-known) solution. Being aware of these different focuses is essential for building on findings based on the Einstellung effect.

3.3.1. Impacts

As already mentioned in chapter 1, Einstellung affects everyone – although to different degrees, as can be seen in subsection 3.3.2. The tendency to start from a point of familiarity immediately leads to the risk of focusing all attention on previously applied strategies; thus, they prevent new and possibly better ideas to come to mind. There is some controversy about whether this strongly negative learning transfer is a phenomenon that only occurs in experimental environments (Singley, 1989). However, other researchers that have conducted research with and around the Einstellung effect investigated this hypothesis and found indications that the effect extends beyond experimental environments (Woltz et al., 2000; Bilalić et al., 2008b). Especially the mistakes that experts commit are often based on the fact that they incorrectly identify a situation as being familiar to them (Reason, 1990). Moreover, Bilalić et al. (2008b), explicitly pointed out that errors by experts resemble the errors that can be reproduced in experimental environments. This shows that both short- and long-term influences can be affected by the Einstellung effect.
It is so quickly adapted that this inability to escape one’s activated mental set can already occur after a single previously solved problem (Crooks & McNeil, 2009). This singular experience can already trigger a “first thing that comes to mind” effect when another similar situation is encountered. From there on, attention is mainly or even exclusively focused on this known solution strategy, ignoring other alternatives (Bilalić et al., 2008b). Traditionally, this has been shown by conducting highly controlled experiments (Luchins, 1942), where participants encounter training problems and subsequently dissimilar problems that are designed to appear familiar. Performance is significantly reduced for the dissimilar problems, when participants previously learned a certain strategy to solve this familiar appearing set of problems. The high impact of Einstellung is revealed when comparing the results to a control group that only encountered the dissimilar problems. These ‘naïve’ participants find the more optimal strategy much more often, which shows that the problems are not intrinsically difficult (Luchins, 1942; Bilalić et al., 2008b; Ellis & Reingold, 2014).

Building on this point, one may question the overall complexity of the presented problems and neglect the effect’s impact on complex real world problems. While only limited research exists in this field, Woltz et al. (2000) investigated the existence of Einstellung in more complex problems and concluded that it was observable, although with a smaller effect size. Woltz et al. (2000) still point out that its impact could be even more significant, due to the fact that it can actually occur in more complex and possibly more critical systems and environments. They, furthermore, mention that the participants were not aware that they were strongly influenced by their previously applied strategies. This was the first time that a researcher explicitly focused on the participants’ general awareness of being influenced by Einstellung. Previous experiments put more weight on the participants’ inability to detect more efficient strategies (Vallee-Tourangeau et al., 2011; Bilalić et al., 2008b; Sweller et al., 1983). However, already the awareness of a negative learning transfer can be used to improve the situation, while nothing can be changed if a problem occurs without anyone knowing what caused it.

Luchins (1942) himself saw the greatest implications of Einstellung effect in education. He provides numerous examples of school situations, where Einstellung occurs; thus, negative learning transfers are created. One example is the value of speedy responses that are often required in question and answer situations and lead to students often relying on familiar solution strategies. This encourages students to be led by their learned habits instead of thinking about alternative and possibly more efficient strategies.

### 3.3.2. Victims

The Einstellung effect is an almost universal phenomenon, where no strong correlation to certain characteristics of humans seem to exist. Luchins (1942) summarized in his original work from 1942 that he did not find any correlation between the occurrence of Einstellung and the age or the educational level of the participants. Cunningham (1965) later used age, sex, and IQ as possibly correlating factors, but none of them proved to have any effect on the occurrence of the Einstellung effect.
However, one cannot say that everyone is affected or that everyone who is affected, is equally inflexible to change her mind. Luchins (1942) observed that there are individual differences that cannot be generalized. Some participants seem to develop no Einstellung effect at all, some have a higher resistance and others recover faster. One study showed, e.g., that depressive patients with a high suicidal risk “[... ] tend to exhibit a narrowing of perceived options and difficulty in considering alternatives.” (Greenberg et al., 2010, p. 6).

This indicates that there may be cases, where stronger proneness to the Einstellung effect may exist.

It was mentioned in section 3.3 that Luchins (1942) saw the Einstellung effect as something that is created in the experimental situation and not necessarily bound to previous experiences prior to the experiment. Years later, researchers looked specifically into the question of how experience and especially expert knowledge can have an influence on participants developing an Einstellung effect. Wiley (1998) investigated the effects of domain knowledge on problem solving and came to the conclusion that it constrains the generation of strategies and that it could be generalized across many domains and problems. 10 years later, Bilalić et al. (2008a) designed an Einstellung experiment for chess players of all skills. Their results were more twofold than Wiley (1998)’s. On the one hand, yes, experts were equally prone to the Einstellung effect as novice players. On the other hand, the higher the expertise, the less susceptible they were (Bilalić et al., 2008a).

The conclusion that one can draw from this is that, when conducting Einstellung experiments, one may not need to have balanced samples in terms of age, sex, education, or expertise. When it comes to the occurrence of the effect, most people, apart from few exceptions, should be susceptible to it; thus, they act accordingly.

3.3.3. Occurrence

To design an Einstellung experiment, it is necessary to understand the exact causes of what makes the effect appear. Previous sections already mentioned some factors superficially, but this section will go into more detail.

There seem to be four attributes that play a role:

- Familiarity
- Repetition
- Recency
- Success

All of these come together and create the basis for Einstellung to occur.

Familiarity is probably the single most important attribute in this list. As mentioned in the introduction, familiarity is necessary for all learning transfer to occur. Without our ability to identify current problems as being similar to ones we have encountered before, we would create cognitive strain every time we do something. With familiarity,
we can recycle strategies from old problems and use them for new ones. Similarity, hereby, describes the relationship between problems, whereby familiarity describes the ability of a person to detect this relationship – possibly incorrectly. This incorrect identification, created by a sense of familiarity, is then one of the four attributes that facilitates the occurrence of Einstellung (Sweller & Cooper, 1985; Bilalić et al., 2008b, 2010).

Repetition or frequency describes the number of times a participant is confronted with a certain problem. As previously mentioned in section 3.3, there has not been a lot of research concerning this point. However, we can assume that an Einstellung effect can already occur after a single previously encountered problem (Luchins, 1942). With each additional problem, the effect will get stronger; thus, it becomes harder to recover from it (Sweller et al., 1983). The repetitive use of a single solution strategy is an indicator that the participant did indeed recognize a link between the current problem and previous problems of the same kind.

Recency is probably the least strong attribute required to cause Einstellung, as the effect has been shown with participants that had no immediate training, but long-term experiences and knowledge (Bilalić et al., 2008a). However, when one is interested in the performance of novices, the impact of recency can be significant (Luchins, 1942). Wiley (1998) argues that novices can probably recover from the Einstellung effect by simply waiting, which may be due to the fact that this knowledge is not as strongly encoded in their memory as it is with expert participants.

Success, or otherwise described as the history of success (Lovett & Anderson, 1996), is another key component to the Einstellung effect and learning transfer in general. Solving an encountered problem successfully will fix it more firmly in one’s mind and makes it, thereby, easier retrievable for future applications. This is of course connected to repetition, by making it more likely that a strategy will be recalled, the more often it has been successfully used in the past (Lovett & Anderson, 1996). On the other hand, failing to apply a strategy, which has been previously successful, does not seem to be enough to make the participant look for alternative strategies (Bilalić et al., 2008b).

While not all of these attributes necessarily need to be present to create an Einstellung effect, they do play together and facilitate its occurrence.

3.3.4. Applications

Besides testing and investigating the Einstellung effect by itself, it is also possible to use its existence as a vehicle to test other hypotheses that may not aim at uncovering more information about Einstellung. The fixation of the mind to use the first activated operator that comes to one’s mind, due to the similarity between the current and a previous problem, can be described as specific knowledge structures that were developed (Sweller et al., 1983). These may be inappropriate, if the similarity is incorrectly categorized, but they are, nevertheless, structures with a certain longevity. As was mentioned in subsection 3.3.2, these structures are not only a phenomenon that occurs in the direct situation of experimental testing, but can also be observed in experts. That is why the observation of structured and fixed operator selections may be an indicator for participants developing or having developed expertise in the domain that the experiment is based on (Sweller
et al., 1983). This expertise then leads to the faster identification and application of certain operators that are believed to solve the problem at hand, which, if incorrectly identified, in turn correspond to the very effect of Einstellung (Woltz et al., 2000).

However, also the absence of fixed behavior could be an indication of expertise (Sweller & Cooper, 1985), but not in the same way as when the Einstellung effect is present. When a participant is trained to use a certain strategy and that same participant does not use this previously trained strategies to solve subsequent problems, but more effective solution strategies, the conclusions can be twofold. Either the participant may not have perceived a similarity between the problems or she may have had or developed the expertise to be more mindful about alternative strategies.

Colvin et al. (2001) investigated the effect that frontal lobe lesions could have on certain cognitive abilities and utilized the Einstellung effect with water jar tasks as a way to test this. They argued that the original experiment, conducted by Luchins (1942), presents a combination of problem solving skills and planning strategies; thus, is allows to identify specific deficits that their patients might have. Colvin et al. (2001) focused very much on the approach that their participants chose to solve a problem. Which problem solving strategy did they use, for how long and how successful were they? Their conclusions were that participants do not plan ahead when confronted with water jar tasks; thus, they often display a hill-climbing strategy (see subsection 3.2.2). However, strictly following this strategy will not solve the problem, which is why the participants had to recognize the need to violate the strategy in order to solve the problem.

3.4. Microworld

The experiment conducted for this thesis makes use of a microworld to create different Einstellung problems, the critical problem, and to automatically record the interactions of the participants with its various components.

Papert (1993) was the first one to use the term microworld, when talking about simulations with an educational application. It has since been argued that adding the interactivity of a simulation to cognitively demanding tasks offers a much closer representation of real world problem solving behavior, as the actual manipulation includes the problem solver in a reactive environment (Hutchins, 2010). When it comes to designing these cognitive tasks in relation to what is to be accomplished, Brehmer (2005) suggests that a microworld has to become increasingly high-fidelity, the more specific a desired behavior is supposed to be shaped. However, if one is more interested in the training of higher cognitive functions, the simulation may represent the targeted domain in a more abstract manner.

Interacting with microworlds requires skills that are more often observed in the field than in experimental environments (Vallee-Tourangeau et al., 2011), whereby more representativeness can be achieved within a controlled environment. One aspect that reduces the representativeness of microworlds is the fact that each participant is in control of all possible manipulations (Dörner, 1993). This position of total power presents a specific
situation in the real world; thus, any attempt at generalizing the results needs to take this attribute into consideration.

3.4.1. Characteristics

There are three main characteristics that all microworlds have in common (Dörner, 1993; Brehmer, 1996):

- Complex
- Dynamic
- Opaque

Dörner (1993) describes *complexity* as the necessity of participants to consider several things at once – either via directly observable objects or by keeping something in mind. Brehmer (1996) puts this in other words and says that it refers to “many interconnected systems that lead to goal conflicts and side effects.” (Brehmer, 1996, p. 232), which extends Dörner (1993)’s description with possible results that can occur.

*Dynamic* is described very similarly by both researchers, as the microworlds being manipulated by the participant’s actions and autonomously – based on built-in rules and functions. It is, thereby, a continuously evolving system that changes based on its previous states and has determinant future states (Dörner, 1993; Brehmer, 1996). This dynamic characteristic leads to participants having to continuously make decisions in real time, which makes them dependent on the dynamic of the microworld, but also influences it (Brehmer & Allard, 1991). Brehmer (2005) later called this a circular relation between the participant’s behavior and the effect it has on the current and future states of the microworld.

*Opaque* refers to the characteristic that not everything is in plain sight. Some statuses, data, or relations have to be inferred by the participant, which means that she has to search for them and form her own hypothesis about how the system works (Dörner, 1993; Brehmer, 1996).

Dörner (1993) says that the goal that is to be achieved by a participant of a microworld is to either reach a certain final state or to maintain a certain state. While trying to achieve this, the participant is planning, solving problems, and constantly making decisions in a continuous flow of actions that require the coordinated execution of all these activities. The interaction with a microworld cannot, according to Dörner (1993), be described as the isolated usage of certain cognitive abilities, but the organized interplay of several efforts to manipulate the presented objects in order to achieve the goal.

During this process, the participant makes use of the information that is presented to her and evaluates it, in order to create a mental model of how the system works. This step is highly dependent on the observability of the system status and the available actions, as only these can be used by the participant to form a model (Dörner, 1993). The mental model then helps the participant to categorize the information that is observed, predict future events, and deal with the dynamically changing environment – even under stressful
circumstances (Brehmer, 2005). However, Dörner (1993) points out that due to the variances that can exist in the mental models and the continuously changing environment for each participant, it is suboptimal to study individual responses. Instead, one should focus on strategies that can be identified by looking at the patterns of individual responses and how these change over time.

As a microworld is a complex, dynamic, and opaque environment, the participant will try to minimize the cognitive strain that it has to invest in order to solve immediate problems. This can lead to a number of errors, which are relevant when one is interested in analyzing the participant’s behavior and strategies. Dörner (1993) compiled a list over nine errors, whereof three are of particular interest:

1. People tend to adopt an ad hoc behavior that only considers demands of the situation at hand.

2. There is a tendency – especially in crisis situations – to overlook checking the effects of the courses of actions taken. The participants take a ballistic attitude and fire the actions without checking on their effects.

3. When the problems become overwhelming, the participants show ‘thematic vagabonding’ or ‘encystment’ (i.e., they shift between goals or concentrate upon one small goal that can be achieved, ignoring other aspects of the problem that must also be solved).

The first error describes both the chance of an occurring Einstellung effect and the use of hill-climbing to advance forward without causing too much cognitive strain.

The second error can be connected to the mental models that participants develop when encountering a microworld. Without a specific and correct understanding of how one’s actions will affect the current status, the participant is likely to continue by a trial-and-error approach.

The third error is something that can be strongly connected to the strategy that individual participants are using. Due to the dynamic and constantly changing structure of a microworld, strategies that focus on certain components, without taking the whole environment into account, can lead to inferior results.

3.4.2. Microworld usage

Dörner (1993) and Brehmer (2005) both name two lines of research that have used microworlds: Individual differences approach and the experimental or system characteristics approach.

The individual differences approach gathers the data of numerous participants, while these are interacting with a microworld. The goal is later to identify differences between the performance and behavior of individual participants, in order to predict the performance of participants in general, based on these identified differences. These aspects solely come from the participant, as the microworlds and previously given instructions were the same for all participants (Dörner, 1993).
The *experimental* or *system characteristics approach* on the other hand looks explicitly at independent variables that are not based on participants’ individual differences. As the name already suggests, this approach is similar to traditional psychological experiments, where participants are assigned to different conditions that represent different microworlds or instructions. Furthermore, as with other psychological experiments, the results are only as stable and reproducible as the level of data abstraction and reduction allow them to be. This can include the focus on whether certain goals were achieved and what strategies were used – instead of looking at individual decisions that were made during the experiment (Dörner, 1993).

As mentioned in section 3.4, microworlds, although studied within laboratories and with an *experimental approach*, are quite realistic representations of the real world; therefore, they bring elements of field research into the experiment. These microworlds are not necessarily as complex as real world systems, but already a slight increase in complexity allows for the observation of more realistic and complex cognitive mechanisms (Dörner, 1993).

Dörner (1993) and Brehmer (2005) discuss in their papers a variety of different aspects that can be studied with the help of microworlds. Some of these include the following:

- How do participants manually control systems and what are the demands that control tasks impose on them?
- How do participants handle the feedback, complexity, and how does their behavior unfold over time under controlled circumstances?
- What models and intentions do the participants develop in their interaction with complex systems and how?

This thesis focuses on the last point, the development of mental models and additionally, if these mental models can change during their time with the microworld. It is a question of how participants make sense of this complex and dynamic environment, without having all information and only limited time to analyze the situation and solve the problem.

The computerized implementation of microworlds makes it possible to log each and every action that the participant makes. Due to this, one can identify patterns – if they exist – that can represent the strategies that the participant is following, and detect when strategies are changed. This can then be used to control if an Einstellung effect did occur, by comparing the behavior patterns from the training scenarios to the patterns in the following problems.

In general, the usefulness of microworlds is based on the combination of its characteristics. On the one hand, its complex and dynamic nature allows for the observation of more real world oriented cognitive mechanisms and their development. On the other hand, the ability to automatically log every action makes it afterwards easier to handle and analyze the data in an efficient and flexible way.
4. Method

People tend to transfer their knowledge from previously solved problems to new problems that they identify as being the same problem as before. While this learning transfer is mostly positive, it can quickly turn into a negative transfer, if the encountered problem is incorrectly identified as being solvable with the previously applied strategy. Thus, the new problem may be less efficiently solved or not at all, only because of the habitual application of the wrong strategy. This Einstellung effect can, however, also be used to identify if people use previously applied strategies when encountering new problems. This makes it possible to, subsequently, compare the performance of these strategies. Studying this is the purpose and basis for the following experiment.

4.1. Hypothesis

This thesis explores the differences in performance for two groups that were trained on two different strategies, when subsequently encountering the same problem. One strategy, from here on referred to as type I, takes an intuitive approach of solving a problem – by training people to solve easier problems of the same kind. The other strategy, from here on referred to as type II, breaks down the problem into its most basic component and trains the participants on solving this isolated component.

The hypothesis is that the participants who receive the instructions and goals for the condition type II will solve the subsequent critical problem faster than participants who receive the instructions and goals for the condition type I.

4.2. Design

The experiment follows a between-subjects design with two conditions, type I and type II. Each condition represents a different strategy that is connected to a different set of instructions and goals. These are represented in a microworld that provides the basis for learning, executing, and evaluating both conditions. An overview of the experimental design is presented in Figure 4.1.

The training phase of both participant groups has a duration of 15 minutes. This enables the participants to complete their tasks several times, while simultaneously introducing a slight stress factor. The idea is that this stress will lead to the participants using their first successful strategy immediately again, without thinking about alternative ways of solving the problem. The more often the participants complete their goal during the training phase, the more likely it is that they will also use their strategy for
the critical problem. Completing the goal will result in a point, which is added to a counter that will subsequently be called the score.

The critical problem has a time limit of 60 minutes and can only be completed once. When a participant reaches her goal, she will immediately advance to the expert problem. If a participant has not reached her goal within these 60 minutes, she will continue to the expert problem, nevertheless.

The expert problem is similar to the critical problem, apart from it being harder to solve and not having a limit on how many points the participant can gather.

Table 4.1 shows the data that is collected for each participant. The participant related data is gathered via an anonymous survey prior to the test (see appendix A), while the dependent variable is gathered automatically from the microworld. Other, possibly confounding, variables are the two different rooms that participants solve the tasks in and the time of day they start the experiment.

<table>
<thead>
<tr>
<th>Dependent variable</th>
<th>Participant data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Task time</td>
<td>Gender</td>
</tr>
<tr>
<td></td>
<td>Age</td>
</tr>
<tr>
<td></td>
<td>Country</td>
</tr>
<tr>
<td></td>
<td>Study program</td>
</tr>
</tbody>
</table>

Table 4.1.: Data collected during the experiment

The two experimental conditions type I and type II have one dependent variable, which is the time it takes to solve the critical problem (task time). The success criterion, which needs to be fulfilled in order to solve the problem, is communicated to the participants through instructions that are given to them directly before encountering the microworld.
The participants were until the end of the experiment not aware that different conditions existed. They, furthermore, did not see, read, or hear the term that was given to their condition; in this case type I and type II.

### 4.2.1. Conditions

Type I represents the condition that views problems as a whole and also tries to solve them from that perspective. It is, thus, that the critical problem is more similar to this condition, in order to replicate the Einstellung effect as it has been traditionally studied. Building this analogy between problems, but making the training sequences easier to solve, reduces the importance to master the critical aspects.

Type II represents the atomic view, which focuses on training for the most critical part of a problem that has to be solved in order to solve the whole problem. The goal is, thereby, to facilitate tactical learning that is aimed towards solving exactly this specific problem component. While this problem is less similar to the critical problem in terms of how the goal is described, the underlying strategy of solving the problem is according to the hypothesis of this thesis very similar to the one that is used during the type II training.

### 4.3. Microworld

Figure 4.2 shows an annotated version of the microworld, which was used to conduct this experiment. This is the visual representation of the microworld that every participant interacted with. Representing the different tasks from section 3.3, there are different goals set for the participant, depending on the stage in the experiment (see section 4.2) and the participant’s condition (see subsection 4.2.1). The microworld was created using HTML, JavaScript and Scalable Vector Graphics.

![Annotated screenshot of the microworld](image)

**Figure 4.2.: Annotated screenshot of the microworld**

Figure 4.2 shows five tanks that are connected to each other by pipes. Each tank has a fill level, which is represented by a dark blue area in that tank. Moreover, each tank has a valve connected to it, which is represented by a slider with a range of six values. It can be manipulated by the user to adjust the water flow of that specific tank and, thus, other tanks, as the microworld represents a closed system. The sliders have six operational levels, in which the outgoing water flow increases, the higher the slider level
is set. Due to its closed system property, an increased outgoing water flow of one tank will simultaneously increase the incoming water flow of other tanks. The fill levels are continuously updated every 100 ms, which makes it possible to observe the water flow of all tanks simultaneously.

The previously mentioned elements and mechanics are the same in all conditions and for all participants. This is due to the necessity that participants need to perceive their different tasks as being similar to each other. As described in subsection 3.3.3, this is one of the main attributes problems need to have in order for the Einstellung effect to occur.

Only the target area (see Figure 4.2) changes in the expert problem. It consists of the inner area, which has a dark green color, and the outer area, which has a light green color. Both areas are the same for both conditions, the training problems, and critical problem. The difference is in the expert problem, where the dark green inner area is 60% smaller for all participants.

Furthermore, the microworld displays a countdown timer that depends on the step that the participants are currently in. On the top right side is a score keeper that counts up whenever a participant successfully completes her current task; therefore, she gains a point.

The task descriptions that the participants in the type I and type II conditions received for their Einstellung problems are available in section D.2 and section D.3.

Before learning about the tasks, both groups received an identical description of the microworld, its elements, and how the mechanics work. Moreover, they were provided with a minimal example of the microworld, which included two tanks and allowed them to manipulate the sliders in order to see that the water flow is changing (see section D.1).

4.4. Procedure

Figure 4.3 provides an overview with all steps of the experiment.

The ambition was to test all participants as simultaneously as possible. Due to the limited capacities of Uppsala University's computer rooms, two nearly identical rooms were booked that, together, could provide space for half of the participants and ensure that each participant will have one empty desk between her and the next participant. Detailed information concerning the localities are provided in section 4.7.

The experiment was conducted in two waves directly after each other because both rooms, together, could only provide space for half the participants. The first wave started at 15:30 p.m. and the second wave 1.5 hours later at 17:00 p.m.. Table 4.2 shows the distribution of the 33 participants over the rooms, time slots, and conditions. Both waves were directly after another to prevent the participants of both waves from talking to each other. No participant left the rooms during the time of the study.
Figure 4.3.: Procedure of the experiment

<table>
<thead>
<tr>
<th></th>
<th>15:30</th>
<th>17:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>Room 1</td>
<td>I 12</td>
<td>II 9</td>
</tr>
<tr>
<td>Room 2</td>
<td>II 7</td>
<td>II 5</td>
</tr>
</tbody>
</table>

Table 4.2.: Participants’ distribution over the conditions, times, and rooms
4.5. Pilot tests

Six pilot tests were conducted in order to remove ambiguity in the provided texts, to test the understanding of the different tasks and to gather data as to what effect size can be expected. All pilot tests were conducted on a MacBook Air (model number 2632) with a diagonal of 13.3 inches, a 1,440 x 900 resolution LCD panel, and an external optical mouse. The study environments were exclusively booked meeting rooms, which ensured a silent environment, without any disturbances and only the participant and the experimenter being present.

The procedure of the pilot tests was due to organizational reasons in two points different from the procedure of the experiment. First, the pilot tests were conducted individually, with the experimenter sitting behind the participant to take notes. Second, the pilot tests did not encounter the expert problem, which means that the session was finished when the participant solved the critical problem or when the time limit of 60 minutes was exceeded. The participants were not informed about the former criterion that finishes the test.

While the pilot tests did provide information that could be used to improve the instructions, they also created the first data sets to estimate if the hypothesis holds. A discussion about the relevance of these preliminary results is included in chapter 6.

4.6. Participants

33 participants were recruited by handing out flyers to students in Uppsala University’s Department of Information Technology at Polacksbacken. The flyer informed them about the general area of the study, without going into detail about the specific hypotheses. A cinema ticket was given to each participant as a reward for participating in the experiment. All participants are novice users, which means that they have not previously encountered this microworld and the tasks connected to it before. The participants were randomly assigned to the conditions and rooms. Detailed information about the recruited participants can be found in appendix B.

4.7. Materials

The outline of the two rooms that the experiment was conducted in, can be seen in Figure 4.4. Both rooms were equipped with 18 computers and mice. The keyboards were moved out of reach and the Firefox Web browsers that presented the microworlds were started in a kiosk mode (Heinaaro, 2011). These measures were included to prevent participants from jeopardizing the experiment and to reduce distractions from visual elements like the task bar and browser frame.

The main relevant hardware difference between these two rooms is the display size of the available monitors. One room (from here on referred to as room 1) was equipped with HP Compaq Elite 8300 All-in-One Desktop PCs, which had a diagonal of 23 inches and a 1,920 x 1,080 resolution LCD panel. The other room (from here on referred to as room
2) was equipped with Samsung SyncMaster 2494HM monitors, which had a diagonal of 24 inches and also a 1,920 x 1,080 resolution LCD panel. Both rooms were well-lit by electrical light during the whole duration of the experiment.
5. Results

A table with the detailed results of all participants can be found in appendix C.

5.1. Scores

Figure 5.1 shows the distribution of scores that both conditions received during the 15 minutes training phase. It is noteworthy that there are six participants in the type II condition that scored zero points, but only one participant from the type I condition that did not score a point. This can have a negative influence on creating the Einstellung effect, as mentioned in section 3.3.

![Figure 5.1: Plot of the distribution of the scores during the training phase](image)

Figure 5.2 shows more detailed information concerning the distribution of the scores. Type I achieved a higher score mean than type II (4.5 compared to 3.3), although type II has one extreme outlier that has a score of 27.

5.2. Critical times

Figure 5.3 shows two histograms representing the time it took both conditions to solve the critical problem (task time). Applying the logarithm to the times creates a close to normal distribution of the results. Three participants from the type II condition did not manage to solve the critical problem within the 60 minutes time frame.

Figure 5.4 shows a box plot, which provides a direct comparison of the task times of both conditions. Outliers are in the box plots, of this thesis, defined as being more than 1.5 times the interquartile range above the upper quartile or below the lower quartile and they are indicated by a black dot. This makes it clear that the participants in the
Critical times Results

Figure 5.2.: Box plot that shows the scores both conditions received during the training phase

Figure 5.3.: Plots of the distribution of the task time (a) and log(task time) (b)
Expert times

Type I condition were on average faster in solving the critical problem (7.9 compared to 11.3 minutes).

![Box plot showing task times for both conditions](image)

Figure 5.4: Box plot that shows the task times for both conditions.

Table 5.1 shows the results of Welch’s t-test applied to the logarithm of the times of both conditions to complete the critical problem. This shows that the different training strategies have either no or only a very small effect, when comparing the task times. Welch’s t-test was chosen for the analysis due to its increased reliability for unequal variances and unequal sample sizes (Ruxton, 2006).

<table>
<thead>
<tr>
<th>Test</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Welch’s t-test:</td>
<td>$t(25.76) = -1.36, p = .187, d = NA$</td>
</tr>
</tbody>
</table>

Table 5.1: Results of a Welch’s t-test on the log(task times)

5.3. Expert times

Figure 5.5 shows two histograms representing the time it took both conditions to receive their first expert point. Applying the logarithm to the times creates a close to normal distribution of the results. Six participants from the type II condition did not manage to achieve a single expert point. This includes the three participants that previously did not succeed in solving the critical problem.

Only the first expert point is considered because the participants did not have the same time to gather points. The remaining data concerning the expert points will be used and analyzed in chapter 6.

Due to the six participants who either did not reach the expert task or have not succeeded in achieving a point, the median and median absolute deviation of both groups will be calculated instead of their means and standard deviations. Figure 5.6 shows that the participants in the type I condition had a median of 8.3, while the participants in the type I condition achieved a median of 11.1. However, the graph hints at participants from the type I condition having a higher absolute deviation from the median than participants from the type II condition (6.1 compared to 4.2).
Figure 5.5.: Plots of the distribution of the expert times (a) and log(expert times) (b)

Figure 5.6.: Box plot that shows the expert times for both conditions
5.4. Instruction problems

By building the instructions into the microworld, the goal was to reduce instruction differences between groups. However, several situations occurred, where the participants have either not read the instructions completely or forgot parts of them. Two individuals explicitly asked the present experimenter what the task is, when they were supposed to solve the critical problem. They did this after trying to solve the goal they thought they were supposed to solve, but that did not turn out to give them points. These individuals were then informed about their goal, but it was not possible to connect these individuals to the data that was collected. Other cases will be analyzed in chapter 6, where the data suggests that multiple participants did either not know what their goal was or on purpose tried to solve another goal.

Moreover, there was one participant that shortly before the end of the experiment asked why the system would not let her score. The problem was that she did not distinguish between having to bring all fill levels into the dark green area or into the light green area. The participant was then informed by the experimenter about the difference of the areas and what her current task is.

All of these questions came from participants in the type II condition.
6. Analysis

There does not seem to exist a strong correlation between the condition and the performance during the critical problem. However, this analysis shall focus on other factors that may have influenced the outcome or could provide insights for future work.

6.1. Pilot data

Before looking at the data from the pilot tests, it is necessary to keep in mind that the surrounding factors were very different from the later conducted experiment. Table 6.1 presents the main differences.

<table>
<thead>
<tr>
<th></th>
<th>Pilot test</th>
<th>Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>Display size</td>
<td>13.3 inches</td>
<td>23 and 24 inches</td>
</tr>
<tr>
<td>Display resolution</td>
<td>1,440 x 900 px</td>
<td>1,920 x 1,020 px</td>
</tr>
<tr>
<td>Individual or in a group</td>
<td>Individual</td>
<td>Group</td>
</tr>
<tr>
<td>Individual observation</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Expert problem</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>Total time</td>
<td>Between 25 and 70 minutes</td>
<td>90 minutes</td>
</tr>
</tbody>
</table>

Table 6.1.: Differences in surrounding factors from the pilot tests to the experiment

Table 6.2 shows the performance of the six pilot tests on the critical problem. Without any further analysis it would already seem as if the effect size is substantially large. However, the values differ strongly from the mean times that the groups achieved in the experiment: 7.9 minutes for Type I and 11.3 minutes for Type II. Due to the large amount of differences between the pilot tests and the experiment, it is impossible to point out one or multiple factors that could have created this difference.

<table>
<thead>
<tr>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>49.51</td>
<td>13.76</td>
</tr>
<tr>
<td>26.02</td>
<td>5.05</td>
</tr>
<tr>
<td>10.15</td>
<td></td>
</tr>
<tr>
<td></td>
<td>22.75</td>
</tr>
</tbody>
</table>

Table 6.2.: Overview of the task times in minutes from the pilot tests

One factor that is assumed to especially have an influence on the creation of mental models is the display size. Due to the automatic scaling of the microworld, the tanks, pipes, and valves are enlarged when the display size increases. This results in an increased
visibility of the system status, which, as discussed in subsection 3.2.3, facilitates the creation of mental models.

6.2. Training score

As discussed in section 3.3, the Einstellung effect can occur after a single application of a strategy. Therefore, this application has to result in a successful outcome that led to solving the problem at hand. Without an initial success, the strategy has a lower probability of being strengthened and, thereby, used in subsequent problems.

When looking at the problem for the type II condition, the strategy is to identify the exact flow relations between the tanks and then find the correct slider settings to achieve a balanced state. However, it is technically possible for the participants to accidentally set the required settings for a balance state. This is less the case for the type I condition, but it is also there the case that increased success should result in a strengthening of the used strategy. Taking this into account, the following analysis will only look at the data of participants who have achieved a training score of two or higher.

While this does increase the probability that the remaining participants will have incorporated the use of a certain strategy (see subsection 3.3.3), it also means that over half of the participants from the type II condition did not achieve this state. Six participants received no point at all and four other participants got only one point, which could be attributed to luck. This strong discrepancy could be attributed to the intended strategy and it being too hard to learn within a 15 minute training phase.

Figure 6.1 shows the distribution of times that the participants with more than one training score achieved. As was visible in section 5.2, the participants in the type I condition have a higher variance than those of the type II condition, but the mean value is much more even across both conditions – type I: 5.39 and type II: 5.25.

![Figure 6.1: Box plot that shows the task times of participants with a training score higher than one](image)

As is presented in Table 6.3, the Welch’s t-test on this data subset does not reveal a significant different.
Due to the similarity between the training problems and the critical problem, it can be assumed that a higher score during the training phase will lead to better performance during the critical problem, i.e., less time will be required to solve the problem. An ANCOVA with the task time as the dependent variable, the condition as a fixed factor, and the score as a covariate will reveal if, by taking the covarying effect into account, the condition has an influence on the task time.

Table 6.4 displays this ANCOVA with all participants who solved the critical problem. One participant achieved a training score of 27, which results in a Cook’s distance of 12.7 (Cook, 1977). According to Cook and Weisberg (1982), all values with a Cook’s distance greater than one are highly influential points, which will distort the regression analysis and should, therefore, be excluded from the analysis. In this data set, only the above mentioned participant has a Cook’s distance greater than one and will, thus, be excluded. The ANCOVA indicates that a higher score seems to increase the likelihood of solving the critical problem in a shorter amount of time, but even taking this covariation into account, the condition still does not have a significant influence on the task time.

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>1</td>
<td>841.96</td>
<td>841.96</td>
<td>16.58</td>
<td>0.0004</td>
</tr>
<tr>
<td>condition</td>
<td>1</td>
<td>24.02</td>
<td>24.02</td>
<td>0.47</td>
<td>0.4977</td>
</tr>
<tr>
<td>Residuals</td>
<td>26</td>
<td>1320.34</td>
<td>50.78</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.4: ANCOVA with the task time as the dependent variable, the condition as a fixed factor, and the score as a covariate

### 6.3. Room and time effects

As described in section 4.4, the sample was split up across two rooms and starting times. Although the rooms were almost identical and the starting times for both groups were only 1.5 hours apart, an ANOVA for those two factors (see Table 6.5) shows that the starting time of the participants could have had an influence on their task time during the critical problem.

For this analysis, the three participants who did not solve the critical task were not excluded, but their task time was instead set to 60 minutes, which is the upper limit for solving the critical task. This was done in order to investigate the performance differences that may have occurred due to possibly confounding variables.
6.4. Strategy analysis

As the five sliders (or valves) are the main points of interaction for the participants, it is mainly through those that their strategy can be derived (see section 3.4). This direct kind of observation looks specifically at all settings of the sliders and how they change over time. This is very much based on the six different values that each slider can have. The higher the value – which is represented by moving the slider to the right (see section D.1) – causes an increased water flow out of the tank, which is connected to that slider.

As previously done in section 6.2, only the participants who achieved more than one training score will be considered for the analysis. This is done in order to increase the probability that they actually developed and fixated on a strategy.

6.4.1. Training

Figure 6.2 provides a representation of the exact manipulations for all sliders and participants in the type I condition that achieved more than one training point during the last three minutes of the training phase. Each color represents one of the sliders in the microworld that the participant could manipulate. The y-axis indicates the values that each slider can have at a specific point in time.

Although the amounts of points are very similar (mean: 5.29; standard deviation: 1.33), the graphs suggest that the participants had very different approaches to solving the problem.

While some participants in the type I condition, e.g., P8, P21, and P29, mainly set the sliders to either 0 or 5 (bang-bang control (Artstein, 1980)), other participants in that condition, e.g., P4, P5, and P9 made more use of the values between 0 and 5. However, as the final scores were quite similar, it can be assumed that both strategies were equally successful in solving the Einstellung problem of the type I condition.

Figure 6.3 provides a representation of the exact manipulations for all sliders and participants in the type II condition that achieved more than one training point during the last three minutes of the training phase. In comparison to the type I participants, the amounts of points varied much more (mean: 8.17; standard deviation: 9.41).

The participants P11 and P17 show a clear understanding of how to achieve a balanced state because they repeatedly return to the same state after the microworld is reset, due to them achieving their goal. Apart from P6, the other participants rarely set the sliders on the max value and instead adjusted them mostly between the values 0 and 2.

<table>
<thead>
<tr>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>room</td>
<td>1</td>
<td>0.47</td>
<td>0.47</td>
<td>0.65</td>
</tr>
<tr>
<td>start</td>
<td>1</td>
<td>7.36</td>
<td>7.36</td>
<td>10.32</td>
</tr>
<tr>
<td>room:start</td>
<td>1</td>
<td>2.62</td>
<td>2.62</td>
<td>3.67</td>
</tr>
<tr>
<td>Residuals</td>
<td>29</td>
<td>20.69</td>
<td>0.71</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.5: ANOVA with the factors ‘start’ and ‘room’ on the logarithm of the task times
Figure 6.2.: Plots of all value changes in the sliders for all participants with a training score greater than one in the type I condition during the last three minutes of the training phase.

Figure 6.3.: Plots of all value changes in the sliders for all participants with a training score greater than one in the type II condition during the last three minutes of the training phase.
As all these graphs show the last three minutes of the training phase, we can expect that they represent the strategies that the participants fixed their minds on; thus, showing the strategies they will use for solving the critical problem.

### 6.4.2. Critical problem

Figure 6.4 shows the slider movements of the same type I participants from the training phase during the first five minutes of the critical problem. Taking a look at the same individuals from the training phase, who displayed particularly visible strategies, it is apparent that they used the same approach to the critical problem.

These were P8, P21, and P29 for the bang-bang control, while P4, P5, and P9, made more use of the values between 0 and 5. As it turns out, all three participants who made use of the middle values solved the critical problem faster than the three with the bang-bang control.

Figure 6.5 shows the slider movements of the same type II participants from the training phase during the first five minutes of the critical problem.

In comparison to the type I participants, the separation is not as clear cut for the type II participants. One of the previously mentioned participants, who clearly knew how to reach the balanced state (P11), did solve the critical problem in under 4.5 minutes. However, the participant with a score of 27 from the training phase shows clear indications of a bang-bang control strategy and required seven minutes to solve the critical problem.
Another observation that one can make is how the participants with fewer interactions (P6, P7, and P11) actually achieved the solution faster than the other three participants, which seem to have switched more often between the extreme states of zero and five.

6.5. Expert problem

In comparison to the critical problem, it is in the expert problem possible to gather more than just one point. However, the expert problem did not have a fixed time limit, leading to participants spending different amounts of time trying to solve the problem and gathering points. The following analysis will, therefore, only look at the expert points per minute that the participants achieved during the expert problem. Figure 6.6 shows the distribution of the expert score per minute ratio across both conditions.

Both conditions have almost the exact same mean value and standard deviation (see Table 6.6), which indicates performance differences in comparison to the critical problem. While the type I condition on average appeared to solve the critical problem faster (7.9 compared to 11.3 minutes), the differences are negligible for the expert problem.

An ANOVA with the factors training ‘score’ and ‘condition’ (see Table 6.7) also shows that there is little influence from the condition on how good the participants were in achieving points in the expert problem. On the other hand, it seems that a higher training score does not only result in a better task time during the critical problem, but also a
Instruction problems

Analysis

Figure 6.6.: Box plot that shows the expert score per minute ratio for both conditions

<table>
<thead>
<tr>
<th></th>
<th>Type I</th>
<th>Type II</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>0.187</td>
<td>0.193</td>
</tr>
<tr>
<td>SD</td>
<td>0.153</td>
<td>0.155</td>
</tr>
</tbody>
</table>

Table 6.6.: Mean and standard deviation of the expert score per minute ratio for both conditions

higher expert score per minute ratio during the expert problem. As the performances are quite equal among both conditions, it could be assumed that the participants from the type I condition were slowed down due to their lack in training of finding a balanced state.

<table>
<thead>
<tr>
<th></th>
<th>Df</th>
<th>Sum Sq</th>
<th>Mean Sq</th>
<th>F value</th>
<th>Pr(&gt;F)</th>
</tr>
</thead>
<tbody>
<tr>
<td>score</td>
<td>1</td>
<td>0.10</td>
<td>0.10</td>
<td>5.62</td>
<td>0.0265</td>
</tr>
<tr>
<td>condition</td>
<td>1</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.9983</td>
</tr>
<tr>
<td>score:condition</td>
<td>1</td>
<td>0.07</td>
<td>0.07</td>
<td>3.58</td>
<td>0.0713</td>
</tr>
<tr>
<td>Residuals</td>
<td>23</td>
<td>0.42</td>
<td>0.02</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 6.7.: ANOVA with the factors ‘score’ and ‘condition’ on the expert score per minute ratio

However, as some participants gathered points in the expert problem for close to an hour, it is possible that they got tired of the task and their performance, therefore, declined.

6.6. Instruction problems

When it comes to instructions, there appear to have been two main effects that influenced the participants’ performance:

- Participants forgetting the instructions
- Participants misunderstanding them

As earlier mentioned in section 5.4 and discussed in section 6.4, some participants probably did not know what their goal was; therefore, they showed different behavioral patterns than the other participants. This can be seen in Figure 6.7, where the plots show the total deviation of all fill levels of all tanks for each participant of the type II condition in the training phase to the target area of the tanks. Thus, the lower the y-value of the graph, the closer that participant is to solving the type I Einstellung problem. For the goal of the type I condition, the fill levels of the tanks are largely irrelevant, which is why the plots of participants like P7, P17, P11, and P14 display behavioral patterns that were to be expected. However, other participants like P6, P15, P16, and P31 clearly seem to try to solve the type II problem, which was never presented to them and is not the goal that solves their type I problem. There were more occurrences of these patterns than were verbally reported to the experimenters. This could be due to the participants being unaware of their mistake or them being unwilling to express and ask about their progress.

Figure 6.7.: Plots that show the total deviation of the fill levels of all tanks from the type II condition during the training phase to the middle point of the tanks - the target area

One way this could be mitigated is by providing an option to look up the instructions, after the task has already been started. The experiment design then needs to take these timing differences into account that are created when people interrupt their current task.

Another option could be to provide more visual information in the instructions, e.g., in the form of a video that shows exactly how it has to look like for the participant to successfully complete the problem.
Apart from these directly observed effects of forgetting or misunderstanding the instructions, there could be another influence that has to do with how long and carefully the participants read the instructions. The instructions were written to be precise and concise, which means that only skimming through them could result in an inferior understanding of what the following task is about. Therefore, recording the time that the participants spend reading the instructions could, during the analysis, help to identify individuals that lack behind in performance – not due to their inferior strategy, but due to their inferior understanding of the task.
7. Conclusion

The theory behind the Einstellung effect and a qualitative analysis of the results from this study indicate that it could be possible to train people to use predetermined strategies on subsequent problems in the same environment; regardless of whether these strategies are the most efficient ones or not, which reveals an Einstellung effect.

Additionally, when providing them with a specific goal and measuring performance factors, these strategies can be evaluated against each other to investigate if one of them is more efficient or effective than the other. This is where microworlds appear to provide a viable experimental environment, due to their characteristics of having goals, being complex enough to require the development of a strategy and allowing to record different interaction points.

However, while some participants showed clear signs of using the same strategy across problems, even more participants displayed behavioral patterns that could not be linked to a specific previously used strategy. While the Einstellung effect is a clearly observable and reproducible bias in water jar experiments, reliably creating the effect across participants in more complex and dynamic environments seems to depend on more environmental factors. The results provide indications of what some of these error inducing factors might be, e.g., time limitations that do not allow for the fixation on a certain strategy, instruction specificity that allows for the application of multiple strategies, and visibility of the system status due to display sizes, which facilitates the creation of a mental model of the system. Misunderstandings and confounding variations can have severe influences on the mental models that people develop, which makes it harder to compare them against each other.

All of these factors are connected to the environment that presents a problem to a person. A microworld was in this thesis the tool to study the influences of these factors and their effects under close to real world contexts; introducing time constraints, hidden system information, and specific training procedures. It allows to record specific interaction points that make it possible to identify patterns of behavior, which can be the base for categorizing certain problem solving strategies.
8. Future work

In order to reduce the amount of unusable data sets – because the desired strategy could not be observed – it seems that more detailed and specific instructions could be beneficial, as well as testing all participants under almost equal conditions. Possible confounding factors were provided in chapter 7.

Furthermore, while a history of success in using a certain strategy seems to reduce the solving time of similar problems, it does not appear to guarantee the usage of that previously successfully applied strategy on subsequent problems. On the one hand, this effect may occur if the success criteria, which are defined to specify the use of a certain strategy, are too simple; thus, they do not require the creation of a detailed mental model to solve them. Providing more difficult problems could mitigate this issue.

On the other hand, it may be that the critical component of that strategy was either incorrectly identified or incorrectly implemented in the problem solving environment. Investigating means to successfully identify the critical components of problem solving strategies would, thus, be one of the next possible steps to continue this line of research.

Another more data oriented focus can lie in the identification and categorization of strategies, based on the data that is gathered through a microworld. It could be that the mental models of people solving the same problem are indeed similar to each other, but that the means of execution vary so distinctively that the underlying strategy may be difficult to recognize.
Bibliography


Appendices
A. Survey
Survey

What gender do you identify most with?
☐ Male
☐ Female

What is your age?
☐ 17 or younger
☐ 18 - 24
☐ 25 - 30
☐ 31 - 40
☐ 41 or older

In which country have you lived most of your life?
☐ Sweden
☐ Other: ____________________________

How many years did you have English as a subject in school?
☐ 5 or fewer
☐ 6 - 10
☐ 11 or more

What is your current study program?
☐ Information Technology (IT)
☐ Computer Science (DV)
☐ Sociotechnical Systems (STS)
☐ Human-Computer Interaction (HCI)
☐ Mathematics (MA)
☐ Only taking individual courses
☐ Currently not studying
☐ Other: ____________________________

How many years did you study in your current program?
☐ 3 or fewer
☐ 4 - 5
☐ 6 or more
☐ Doesn’t apply
☐ Other: ____________________________

Thank you!

Figure A.1.: Demographic survey
B. Participants
<table>
<thead>
<tr>
<th>uId</th>
<th>Gender</th>
<th>Age</th>
<th>Country</th>
<th>Program</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>m</td>
<td>25</td>
<td>Tibet</td>
<td>Political Science</td>
</tr>
<tr>
<td>P2</td>
<td>f</td>
<td>18</td>
<td>Sweden</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P3</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P4</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P5</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P6</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P7</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P8</td>
<td>m</td>
<td>25</td>
<td>Greece</td>
<td>Mathematics</td>
</tr>
<tr>
<td>P9</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P10</td>
<td>m</td>
<td>25</td>
<td>Croatia</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P11</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P12</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P13</td>
<td>f</td>
<td>18</td>
<td>Germany</td>
<td>Mathematics</td>
</tr>
<tr>
<td>P14</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P15</td>
<td>m</td>
<td>25</td>
<td>Greece</td>
<td>Computer Science</td>
</tr>
<tr>
<td>P16</td>
<td>f</td>
<td>18</td>
<td>Sweden</td>
<td>Individual courses</td>
</tr>
<tr>
<td>P17</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P18</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P19</td>
<td>f</td>
<td>18</td>
<td>Netherlands</td>
<td>Astrophysics</td>
</tr>
<tr>
<td>P20</td>
<td>f</td>
<td>18</td>
<td>Albania</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P21</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P22</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P23</td>
<td>m</td>
<td>31</td>
<td>Sweden</td>
<td>Molecular Biotechnology Engineering</td>
</tr>
<tr>
<td>P24</td>
<td>f</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P25</td>
<td>m</td>
<td>18</td>
<td>Netherlands</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P26</td>
<td>m</td>
<td>25</td>
<td>India</td>
<td>Computer Science</td>
</tr>
<tr>
<td>P27</td>
<td>f</td>
<td>18</td>
<td>Sweden</td>
<td>Computer Science</td>
</tr>
<tr>
<td>P28</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P29</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Computer Science</td>
</tr>
<tr>
<td>P30</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Human-Computer Interaction</td>
</tr>
<tr>
<td>P31</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P32</td>
<td>m</td>
<td>18</td>
<td>Sweden</td>
<td>Information Technology</td>
</tr>
<tr>
<td>P33</td>
<td>m</td>
<td>25</td>
<td>Sweden</td>
<td>Technical Physics</td>
</tr>
</tbody>
</table>

Table B.1.: Detailed information about the recruited participants
C. Full results
<table>
<thead>
<tr>
<th>uId</th>
<th>start</th>
<th>room</th>
<th>condition</th>
<th>score</th>
<th>time</th>
<th>expFirst</th>
<th>expScorePerMinute</th>
</tr>
</thead>
<tbody>
<tr>
<td>P1</td>
<td>1700</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>20.54</td>
<td>22.91</td>
<td>0.05</td>
</tr>
<tr>
<td>P2</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>5.45</td>
<td>31.1</td>
<td>0.05</td>
</tr>
<tr>
<td>P3</td>
<td>1530</td>
<td>1</td>
<td>2</td>
<td>5</td>
<td>2.68</td>
<td>4.02</td>
<td>0.41</td>
</tr>
<tr>
<td>P4</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>2.99</td>
<td>5.39</td>
<td>0.07</td>
</tr>
<tr>
<td>P5</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>3.09</td>
<td>8.29</td>
<td>0.41</td>
</tr>
<tr>
<td>P6</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>2.4</td>
<td>4.2</td>
<td>0.49</td>
</tr>
<tr>
<td>P7</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>7</td>
<td>4.19</td>
<td>7.31</td>
<td>0.17</td>
</tr>
<tr>
<td>P8</td>
<td>1700</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>6.35</td>
<td>50.73</td>
<td>0.02</td>
</tr>
<tr>
<td>P9</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>2.55</td>
<td>3.72</td>
<td>0.39</td>
</tr>
<tr>
<td>P10</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>32.11</td>
<td>19.95</td>
<td>0.03</td>
</tr>
<tr>
<td>P11</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>4.29</td>
<td>7.94</td>
<td>0.3</td>
</tr>
<tr>
<td>P12</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>6.91</td>
<td>15.6</td>
<td>0.17</td>
</tr>
<tr>
<td>P13</td>
<td>1700</td>
<td>2</td>
<td>2</td>
<td>6</td>
<td>2.66</td>
<td>8.26</td>
<td>0.37</td>
</tr>
<tr>
<td>P14</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>22.36</td>
<td>9.99</td>
<td>0.03</td>
</tr>
<tr>
<td>P15</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>7.88</td>
<td>12.2</td>
<td>0.24</td>
</tr>
<tr>
<td>P16</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>6.97</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P17</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>27</td>
<td>7</td>
<td>9.4</td>
<td>0.37</td>
</tr>
<tr>
<td>P18</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P19</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>5.17</td>
<td>27.88</td>
<td>0.05</td>
</tr>
<tr>
<td>P20</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>29.18</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P21</td>
<td>1700</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9.02</td>
<td>22.34</td>
<td>0.02</td>
</tr>
<tr>
<td>P22</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>8</td>
<td>2.64</td>
<td>3.91</td>
<td>0.36</td>
</tr>
<tr>
<td>P23</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>4.84</td>
<td>3.2</td>
<td>0.24</td>
</tr>
<tr>
<td>P24</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>8.38</td>
<td>14.4</td>
<td>0.03</td>
</tr>
<tr>
<td>P25</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P26</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>P27</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>30.55</td>
<td></td>
<td></td>
</tr>
<tr>
<td>P28</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3.41</td>
<td>3.12</td>
<td>0.22</td>
</tr>
<tr>
<td>P29</td>
<td>1700</td>
<td>2</td>
<td>1</td>
<td>4</td>
<td>9.71</td>
<td>33.47</td>
<td>0.03</td>
</tr>
<tr>
<td>P30</td>
<td>1700</td>
<td>1</td>
<td>2</td>
<td>4</td>
<td>8.18</td>
<td>15.68</td>
<td>0.08</td>
</tr>
<tr>
<td>P31</td>
<td>1530</td>
<td>2</td>
<td>2</td>
<td>0</td>
<td>13.44</td>
<td>18.29</td>
<td>0.14</td>
</tr>
<tr>
<td>P32</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>12.05</td>
<td>1.81</td>
<td>0.12</td>
</tr>
<tr>
<td>P33</td>
<td>1530</td>
<td>1</td>
<td>1</td>
<td>5</td>
<td>3.61</td>
<td>22.33</td>
<td>0.25</td>
</tr>
</tbody>
</table>

Table C.1.: Detailed information about the participants’ assignment to a condition and their performance
D. Microworld

D.1. General instructions

Instructions
In this test you will adjust the water flow in a number of tanks.
Each tank has a valve attached to it, which you can use to adjust the water flow. The water will then flow into the tank(s) that the pipe with the valve is connected to.
The valve is represented by a slider and can be manipulated in two different ways:
Left click on the position you want the slider to move to. Left click and hold on the slider, drag it to the position you want it to be and then release the mouse button to set the position. The fill level of all tanks will be updated continuously according to the settings of the sliders.
If a tank becomes completely full or empty, the fill level and all valves will be reset to their original positions and you have to start again.

Figure D.1.: Annotated tank explanation

Here is an interactive example of two tanks that are connected to each other. Please manipulate the sliders in order to understand how to use them.

Figure D.2.: Interactive example of how the interaction works
D.2. Type I instructions

Task explanation
Your task is to get the fill level of all tanks simultaneously into the target area (see Figure D.3). The fill levels do not need to be at the same position within the target area. As soon as all fill levels are anywhere inside the target area, the task is immediately successfully completed.

![Target area and fill level](image)

Figure D.3.: Type I goal explanation

Your goal is to complete the task as often as possible within 15 minutes. After you successfully got all fill levels simultaneously into the target areas, the fill levels and sliders will be reset to their original positions and you can start again.

When the 15 minutes are over, you will automatically be brought to the next step.

D.3. Type II instructions

Task explanation
Your task is to adjust the sliders in such a way that the movement of the fill levels in all tanks is completely stopped. Remember that the test will be reset if a tank is completely full or empty, but other than that, it does not matter what the fill levels of the tanks are – as long as no movement is present.

Your goal is to complete the task as often as possible within 15 minutes. After you successfully stopped the movement of all fill levels in all tanks simultaneously, the fill levels and sliders will be reset to their original positions and you can start again.

When the 15 minutes are over, you will automatically be brought to the next step.