Learning Effects In Syllogistic Reasoning

Bachelorarbeit
betreut durch die Abteilung "Cognitive Computation"
am Institut für Informatik unter Leitung von
apl. Prof. Dr. Dr. Marco Ragni

zur
Erlangung des akademischen Grades
"Bachelor of Arts"
der Philologischen und der Philosophischen Fakultät
der Albert-Ludwigs-Universität
Freiburg i. Br.

vorgelegt von

Raffael Veser
aus Weingarten

SS 2018

Bildungswissenschaft und Bildungsmanagement

Contents

Abstract: Learning Effects In Syllogistic Reasoning

Faculty for Educational Science

Bachelor of Arts
by Raffael Veser

This work examines the reasoning process and how it changes over time. In particular the reasoning process for syllogisms is regarded. Syllogisms are logical statements that contain quantifiers like *all, some and no* and consist of premises which can be drawn a conclusion from. An example syllogism could consist of the two premises, *All doctors are humans. All humans are mortal.* From these premises one can logically make the valid conclusion that all doctors are mortal.

The conclusions which people draw from these kind of statements are in the focus of a lot of research and it could be shown that there are big differences between people as well as within people themselves when confronting the same task several times. The latter circumstance has however not been examined sufficiently by research so far. Because of this a study was conducted in which participants had to draw conclusions from syllogistic premises on three occasions. Participants increased their performance consequently for a subset of syllogisms and answered them more correctly and the reaction times decreased consequently which indicates a learning process. Another important finding of this study was that participants became more consistent over time and therefore changed less of their answers in the third testing.

The results suggest that there are not only big differences between people in the use of strategies for solving syllogisms but also in regard to the process of learning that varies between people. Established theories explaining syllogistic reasoning are lacking an explanation for the development of the reasoning process and future research should therefor take these findings into account.

Zusammenfassung

Lerneffekte bei syllogistischem Schlussfolgern

In dieser Arbeit wird syllogistisches Schlussfolgern und dessen veränderung über die Zeit untersucht. Syllogismen bestehen aus Aussagen, so genannten Prämissen, die quantifizierte Aussagen über Subjekte enthalten. Z.B. könnten die beiden Prämissen lauten *Alle Ärzte sind Menschen. Alle Menschen sind sterblich*, woraus logisch korrekt gefolgert werden kann, dass alle Ärzte sterblich sind.

Welche Schlussfolgerungen Personen aus den Prämissen ableiten variiert dabei nicht nur zwischen ihnen, sondern teilweise auch innerhalb der Person, wenn dieselbe Aufgabe mehrmals bearbeitet wird. Da die bisherige Forschung diesen Umstand nicht asureichend untersucht hat, wurde eine Studie durchgeführt, in der Probanden an drei Terminen syllogistische Schlussfolgerungen ziehen sollten. Die Korrektheit der logischen Schlussfolgerungen nahm für einen Teil der Syllogismem stetig zu und die Reaktionszeiten nahmen stetig ab, was einen Lerneffekt durch das mehrfache Bearbeiten der Aufgaben nahe legt. Eine ebenfalls wichtige Erkenntnis war, dass die Probanden in diesem Experiment konsistenter in ihren Antworten wurden, d.h. dass sie sich in der letzten Testung weniger oft umentschieden als zuvor.

Die Ergebnisse legen nahe, dass es nicht nur große Unterschiede in der Auswahl von Strategien zum Lösen von Syllogismen zwischen Personen gibt, sondern auch, dass ein Lernrprozess stattfindet, der von Person zu Person unterschiedlich sein kann. Bisherige Theorien zu syllogistischem Schlussfolgern sollten deshalb in Betracht ziehen, dass Veränderungen im Antwortverhalten der Normalfall sind, und sollten ebenfalls in der Lage sein, diese Veränderungen zu erklären.

Chapter 1

Introduction

If it rains, your friend will take his umbrella before leaving his house. Consider this everyday example for a logical proposition where if a constituted rule, a premise, a consequence necessarily follows. In this case one has to determine whether it is raining to know that they have to take an umbrella. Now imagine that the above rule still applies and furthermore imagine that this friend of yours has taken his umbrella with him. One might tend to infer from this circumstance that it must be raining outside. But this conclusion is not a valid one to draw because the only information we have is unidirectional, the so called *modus ponens* (McGee, 1985) of the form:

$$p \rightarrow q$$

read as *if p then q*. The converse conclusion

$$q \rightarrow p$$

is logically not correct. The only conversion one can make from a logical point of view is to infer from the absence of an umbrella that the weather cannot be rainy, the so called *modus tollens* (Cohen, 1994).

$$\neg q \rightarrow \neg p$$

where $\neg$ is the sign for negating.
Although this might sound like a trivial example research has shown that people make this sort of wrong deduction most of the times (Cosmides, 1989). Now consider another example of a logical proposition:

None of the Gods are Greek.
All Greeks are humans.

Now the question is what inference can be drawn from this so called Aristotelian syllogism. Once more many individuals draw a conclusion that is logically not correct and they infer that no god is human. The only correct inference would however be that some humans are not gods (Ragni, Riesterer, Khemlani & Johnson-Laird, 2018). But nonetheless humans are still able to use logic, to reason and to draw inferences in given situations. Considering syllogistic reasoning more than a dozen theories of how people reason and come to a certain conclusion have been proposed so far but it remains still unclear which of these theories is the most suitable one - are humans constructing mental models of the given information (Johnson-Laird and Steedman 1978), are they guided by properties of the given premises (Woodworth & Sells, 1935) or is the best way to explain reasoning by implementing a complex parameterized computer model (Khemlani & Johnson-Laird, 2013)? But less clear is how people change over the course of time when facing logical tasks several times - something that has been neglected gravely by research so far and has only been examined by very few studies (Johnson-Laird and Steedman, 1978; Bucciarelli & Johnson-Laird, 1999; Lane, Fletcher, Fletcher, 1983). In this work the process of reasoning and mainly its development is examined with syllogistic tasks simply because syllogisms cover a wide range of logical properties and are easy to construct. Furthermore they are researched very well and can be presented in natural language.

Chapter 2

Theoretical Background

2.1 Human Reasoning

Human reasoning can be defined as "The power of the mind to think, understand, and form judgments logically" (Simpson, Weiner, & Oxford University Press, 1989). This ability is something that is inherent to the human mind and has evolved in such a manner that it exceeds every other species on this planet by far. Even primates which are our closest ancestors and are most likely the closest to humans in terms of intelligence achieve roughly the level of a 2 years old child when dealing with complex cognitive tasks (Tomasello, & Herrmann, 2010). This skill has evolved over a long period of time and it is still highly debated why such a complex and energy consuming structure has been developed in the first place. Some hypothesists claim that it is the result of an ever changing environment in which a more complex brain allows for better adaptation whereas other theories propose a social approach, i.e. that the brain developed in such a way mostly because it allowed for complex social interaction (Herrmann, Call, Hernàndez-Lloreda, Hare, & Tomasello, 2007). And we are in fact using logic and reasoning all the time in our everyday life and during social exchanges. Imagine the following scenario: A young couple is making plans for the next day and one of them proposes to have a picnic in the park. The other partner replies: "But it will rain tomorrow. Let us go to the cinema instead." Their partner agrees and finds that a good idea. Multiple instances of reasoning happened here. The two statements *Let us have a picnic* and *But it will rain tomorrow* are only two statements without any connection at first. But by using knowledge of the world, one infers immediately that it is generally not a good idea to have a picnic when it rains, because it is done outside. Thereafter, the other partner also immediately knows why going to the cinema is a good idea. They draw the conclusion that going to the cinema makes sense because it is in the inside of a building which means they won't get wet in the rain. This example illustrates the necessity of logic and the ability to draw conclusions for social interactions of all kinds. The possibility to draw these kinds of conclusions arises from different rules of communication such as proposed by Grice (1978) and theories of

relevance where the addressee assumes the content of the speaker to be of relevance for the context of the discourse (Wilson, 2016).

2.1.1 Logic in Psychological Research

The example illustrated above is a good example for deductive reasoning. In deductive reasoning one draws a logically sound conclusion from given information, e.g. when the information is that it is raining then it is a logically valid conclusion that you will get wet in the rain when being outside. In contrast, when reasoning inductively, drawn inferences are in the form of generalizations. E.g. if one has seen thousands of white swans in their life so far they might infer that all swans are white. If these inferences are drawn explicitly or implicitly is another means of distinction. In psychology, research is mostly restricted to explicit reasoning tasks in which reasoning outcomes can be checked objectively by others to be correct or incorrect (Evans, Newstead, & Byrne, 1993). A lot of psychological studies have focused on reasoning processes and outcomes. One of the most thoroughly studied tasks in this domain is Wason's selection task.

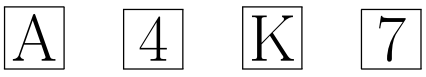

Figure 1: Wason Selection Task
Rule: If there is a vowel on one side, then there is an even number on the other side

The task that participants are given in this experiment is rather simple. The instruction is to turn around as few cards as possible to determine whether the rule is complied or not. The most frequent card which participants turn around, is the card with the letter 'A', which is of course correct and the most forward deductive test. The second most frequent answer however is to turn around the card with the number 4. This might be conclusive at a first glance but from a logical point of view it is incorrect, because by turning the card with the number 4 there is no gain of information, as a consonant on the other side would not violate the rule. The other card that has to be turned around is the one with the number 7 because a vowel on the other side would violate the rule. This conclusion is only drawn by 5% of the subjects, whereas the conclusion to turn around the card with the number 4 is drawn by 45% (Stenning & Lambalgen, 2012). Apparently

when, logically seen, facing the same task embedded in a different scenario people tend to improve their reasoning skills. E.g. an alteration of this task could be as follows:

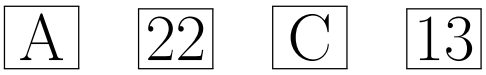

Figure 2: Alternative version of the Wason selection Task

Rule: If a person drinks alcohol (A) then their age on the other side of the card has to be 18 or higher.

Interestingly if participants are again asked to turn around the cards needed to determine whether the rule is followed, 75% of them give the correct answer and turn around the cards A and 13 (Cosmides, 1989). From a pragmatical point of view this makes sense, because there is no need to turn around the card with the number 22 on it. If a person is 18 years or older then they can essentially drink whatever they want and to drink Coke (C) is also allowed for everyone - the observation that the inference one draws is dependent on the context shed new light on the notion of reasoning.

2.1.2 Theories of Human Reasoning

Regarding such results found in Wason's Selection Task questions about how humans reason arise and are still debated. One explanation, called the memory cueing hypothesis, proposes that being able to retrieve past experiences with a problem such as found in the drinking version of the Wason's selection task, facilitates performance. This is supported by an experiment where participants had to solve an alternative version of the Wason Selection Task containing the rule that a letter must have a stamp on it when it is sealed. Since this regulation is commonly known in British countries, people are good in solving this task whereas there is no such restriction in the US and therefore performance differences can be observed (Griggs Cox, 1982). Some hypothesists also claim that logical tasks are not inevitably solved by logic but rather by an attempt to maximize insight and to raise the probability of gaining new information. Incorrect answers would thereby not be the result of failed logic but the application of a heuristic (Chater Oaksford, 1999).

Another distinction that can be made is whether human reasoning is monotonic or non-monotonic. Monotonic reasoning means that the validity of previously drawn conclusions will not be affected if new information becomes available. E.g. if the following two sets of conditionals are given.

(a) If Lisa has an essay to write, she will stay late in the library

(b) If the library stays open, then Lisa will study late in the library.
If Lisa has an essay to write, then she will study late in the library.

The latter conditional leads to a more cautious approach to draw a conclusion. When given the information that Lisa has an essay to write and the additional information in (b), a significant fewer amount of people answer that she will be studying late in the library than those who only have the information of (a). As also can be seen when facing logical tasks the context in which they are embedded seems to influence the applied logic and hence the monotony of human logic seems not justifiable (Ragni, Eichhorn, Bock, Kern-Isberner, & Tse, 2017).
But what are the processes behind human reasoning? In general human cognition can be broken down into the three basic processes of granulation, organization and causation. Granulation describes the process of decomposing the whole into parts and forming granules. This granulation process however results in fuzzy granules for most of the human reasoning. This is an ability guided by an unstable, imprecise environment with partial knowledge and truth which allows for rational choices in these circumstances (Zadeh, 1997). Thus again this can be viewed as an additional remark to the lack of straightforward logical processes in human cognition. Supporting this belief dating back to the 1970s two distinct types of reasoning were introduced to be relevant for describing the human apparatus of reasoning. On one hand there are type I processes that are fast, rely on intuition (which one could also call heuristics) and are mostly unconscious as well as low in cost. Opposed to that there are type II processes which demand a high amount of concentration and are therefore more costly. These type II processes involve analytical thinking and are only active when deliberately chosen (Le & Wartschinski, 2018). So whether a task can be solved logically correctly is a question of how well the two systems are used.
Also the relatively young approach of *quantum cognition* states that human reasoning involves two distinct processes, namely logical and emergent reasoning. Emergence leads to creation of new conceptual structures and logic means the application of probabilistic

rules. Experimental findings such as found in the Wason selection task can be accounted for by conceptual emergence processes (Aerts, Gabora, Sozzo, & Veloz, 2011).
The use of mental models is an additional approach to explain reasoning with the help of mental images. Johnson-Laird (2010) proposes to understand reasoning as a simulation of the available information. Integrating this mental imagery with prior knowledge is what determines the process of reasoning - and not the application of formal logic rules. In this theory the use of counterexamples is a common strategy to assert whether a given statement or a drawn inference can be falsified. E.g. when one is confronted with the following statements

More than half of the people in the room are English
More than half of the people in the room are German
Does it follow that more than half the people in the room are English-Germans?

By constructing a counterexample one can easily determine that a possible configuration could include 6 English, 6 Germans and only 2 English-Germans.
What the different theories about human reasoning have in common, is that they account for possible flaws in the reasoning process and they furthermore state that human reasoning is not driven by rational logic alone and involves some kind of heuristics or generalizations.

2.2 The Syllogism

Dating back to the ancient Greeks, the philosopher Aristotelian introduced what can be called the first rules of logic with this monadic reasoning. Monadic means that a quantifier assigns properties with determiners like "all", "most" and "some" to a set of individuals (Khemlani Johnson-Laird 2013). He introduced assertions containing quantifiers of the form

All A are B
Some A are B
Some A are not B
No A are B

(Stenning Lambalgen, 2012). Until first order logic was introduced in the late 1800s, the Aristotle syllogism was the main instrument used for the logic of quantifiers and is to the current day one of the best researched tasks in logic. The classical scholastic syllogism is arranged from two premises of the above form and an analogical built conclusion following from the two premises. In the case of non-valid syllogisms the only correct conclusion is however that there is no valid conclusion possible. In most cases the subjects of each statement are paired with affiliations like professions. A typical syllogism would therefore look like this:

Some artists are bakers
All bakers are chemists
$\therefore$ *Some artists are chemists*

In this case the *mood* of the conclusion would be affirmative existential. In general four different moods of a statement can be distinguished:

All A are B	*Affirmative universal (abbreviated as "A")*
Some A are B	*Affirmative existential (abbreviated as "I")*
No A are B	*Negative universal (abbreviated as "E")*
Some A are not B	*Negative existential (abbreviated as "O")*

Considering the two premises of a syllogism 4 x 4 different combinations of moods are possible which would result in 16 different types of syllogisms. But there is yet another distinction to be applied to the premises called the *Figure* of a syllogism. Abstractly the subjects of the premises can be referred to as A, B and C, where A is the subject solely found in the first premise, B is the subject that occurs in both premises and C is the subject that only occurs in the second premise. Four different figures can be distinguished:

Figure 1	Figure 2	Figure 3	Figure 4
A-B	B-A	A-B	B-A
B-C	C-B	C-B	B-C

In the example above the figure would therefore look like this.

A-B (artists-bakers)
B-C (bakers-chemists)

Combining each figure with each mood results in 4 x 4 x 4 = 64 different pairs of possible premises. Every syllogism is typically labeled with two letters denoting to the mood of the premises and a number for the figure. So the above example would be labeled IA1. For these 64 possible pairs of premises only 27 can be drawn a logically valid conclusion from (regarding a relationship between A and C, of course there can be drawn conclusions containing only subjects of one premise which however is not the aim of syllogistic reasoning) from the other 37 pairs of premises no logically valid conclusion can be drawn (Khemlani & Johnson-Laird, 2012).

2.2.1 Syllogistic research

The reasoning behind syllogisms has been researched for almost a century now (e.g. Woodworth & Sells, 1935; Wilkins, 1928). The varying difficulty of syllogisms has been established quite early showing that some syllogisms are even feasible for children around the age of 10 who can draw valid conclusions spontaneously whereas other syllogisms seem to be too difficult even for adults (Johnson-Laird, Oakhill, & Bull, 1986). To be more precise, it could be shown that participants for example are experiencing more difficulty facing existential rather than universal premises. Woodworth (1935) suggested that this may be due to the ambiguity of the word *some* which can account for either a subset of X or the whole set of X in logic. In everyday language the conversion

$$\text{Some X are Y} \leftrightarrow \text{Some X is not Y}$$

is made quickly, in the terms of syllogistic reasoning *some X are Y* does however not exclude the possibility that All X are Y.

One of the most robust effects found in syllogistic reasoning is the big amount of individual differences. E.g. participants from an American elite university drew the right conclusions in 55% of all cases whereas participants from an average Italian university drew the right conclusions only on 37% of the problems. Moreover the accuracy in respect to giving a correct answer ranged from 15% to 85%. The higher accuracy observed within the American participants can be explained through correlations between measured intelligence and reasoning abilities (Khemlani Johnson-Laird, 2012) hence the more intelligent a person is, the better they should perform in syllogistic reasoning tasks.

Another phenomenon that can be found is concerning the figure of a syllogism. A figural effect was first described by Johnson-Laird and Steedman (1978), which predicts a certain

bias towards a congruent conclusion of figure. Considering the following example from their experiment this figural bias becomes more tangible:

Figure 1	Figure 2
Some A are B All B are C	All B are A Some C are B
∴ Some A are C (15 subjects) ∴ Some C are A (2 subjects)	∴ Some C are A (16 subjects ∴ Some A are C (1 subject)

As one can see here there is a strong tendency towards conclusions of the form A-C for Figure 1 and for Figure 2 of the form C-A. In this case, both of the conclusions are valid. This bias however also occurs in syllogisms where the converse conclusion is not valid or for invalid ones. In cases where the figural bias contradicts a valid conclusion this syllogism seems to be hardest to solve and, vice versa, easy syllogisms' conclusions coincide with the figure.
For example the syllogism

No A are B
All B are C

is one of the most difficult syllogisms, where most people show the tendency to answer *No A are C* trying to link up the premises' components. The only valid conclusion would however be that *Some C are not A*. In contrast to this, the following can be taken as an example for an easy syllogism.

All B are A
Some C are B

The figural bias leads to *Some C are A* as the most frequent drawn conclusion which is a correct conclusion and therefore shows that an easy syllogism accords with the figure (Khemlani Johnson-Laird 2012).

2.2.2 Theories of syllogistic reasoning

When trying to learn about the internal processes of subjects it becomes apparent that different reasoning strategies are applied. This concerns the premise which they interpret

first, how the interpretation is done and how internal mental models are constructed and used. Hence a big number of theories have emerged trying to explain processes and to predict performance on syllogistic reasoning tasks. These theories mainly diverge into three different basic ideas: heuristic theories, theories based on formal inferences of inference and theories making use of diagrams or sets (Khemlani Johnson-Laird, 2012).

2.2.3 Atmosphere effect

One of the first approaches to structuring the results that syllogistic experiments yielded was proposed by Woodworth (1935). He observed a tendency of participants to draw conclusions that correspond to the mood of a syllogism. Looking back at the four possible moods A, E, I and O of a premise, one can see they contain a quantifier (universal vs. existential) and a polarity (affirmative vs. negative). So in the case of the premise *Some A are B* which would be of the mood I, the quantifier would be existential and the polarity would be positive.
Due to experimental findings, Woodworth suggested the hypotheses that

1. A syllogism that at least contains one premise with an existential quantifier elicits the tendency to draw an existential conclusion and therefore a *some atmosphere.*

2. A syllogism that at least contains a premise with a negative polarity leads to building a negative atmosphere and hence to the tendency of drawing a negative conclusion.

Though debated atmosphere effects seem to be a component of the human reasoning process and emerge in other tasks as well. Revlis (1975) therefore suggested a feature selection model for drawing conclusions trying to outline the mechanisms of deduction. In a first step the features, in this case the quantifier and polarity, are extracted from the premise and then merged into a combined representation. The features from the considered conclusion are then matched with the built up representation. The conclusion is thus accepted when the representations are congruent. It should be noted however that this prediction concerns the most frequent conclusion to be drawn and does not exclude the possibility to draw deviant conclusions.
There are yet some remaining problems with the atmosphere effect, e.g. that the general rules of syllogistic inferences overlap with the principles of the atmosphere effect (Khemlani Johnson-Laird, 2012) but it is nonetheless a theory describing found patterns in research and predicting performance which makes it considerable for syllogistic research.

2.2.4 Mental Models

After discovering a figural effect in syllogistic reasoning, a theory to account for these findings was proposed by Johnson-Laird and Steedman (1978). Their main essence of the theory states that inferences are built up through mental representations that are analogous to the logical properties of a syllogism. The involved processes are described in four steps, namely:

1. Semantic interpretation
2. Initial heuristic combination of the two premises' representations
3. Formulation of a conclusion
4. Logical test of the conclusion

In the first step the premises' properties are represented in a mental image. For the premise *All the artists are beekeepers* one participant from Johnson-Laird and Steedman's study described for example, how he imagined a room full of artists with all of them wearing beekeeper's hats. This mental image could be represented in the form of a list-structure where relations between the objects are outlined. The above example would therefore take on the following form.

$$\begin{array}{ll} artist & beekeeper \\ artist & beekeeper \\ & beekeeper \end{array}$$

In a second step a combination of the two formed models is made from an heuristic approach. This heuristic favors a bias towards conclusions that are linking up the middle-terms of the premises. If the second premise was *Some beekeepers are chemists* the combined representation could look like this:

$$\begin{array}{ll} artist \rightarrow beekeeper \rightarrow & chemist \\ artist \rightarrow beekeeper & chemist \\ & beekeeper \end{array}$$

which would favor the conclusion some artists are chemists.
In the third step this combination of the premises is used to formulate a conclusion. The paths between the items are then used to link them. These paths are in a direction and therefore utter a figural bias. Lastly a logical test is applied to the derived conclusion which tries to falsify it. E.g. one can try to break the link connecting the three entities artist, beekeeper and chemist and thus finding a constellation where the initial conclusion is not fulfilled.

$$\begin{aligned} artist \rightarrow beekeeper \qquad & chemist \\ artist \rightarrow beekeeper & \\ beekeeper \rightarrow chemist & \end{aligned}$$

In this case no artist is a chemist but the constellation does not violate the premises which is why the initial drawn conclusion can be discarded and one could therefore come to the conclusion that nothing follows from the premises (s. Johnson-Laird & Steedman 1978, Bucciarelli & Johnson-Laird, 1999).

2.2.5 Probability Heuristics Model

An approach to shed new light on the notion of reasoning is the Probability Heuristics Model trying to explain reasoning in everyday life and its application onto laboratory scenarios (Chater & Oaksford, 1999). Furthermore it tries to account for incorrect deductions not through a flawed logical process but rather by the invocation of probabilistic reasoning schemata.
Thus, the mismatch between logic and performance is resolved not by invoking flawed logical algorithms, but by adopting a probabilistic rather than a logical computational level theory. One major categorization made in this theory is the informativeness of statements, e.g. *All A are B* is more informative than *Some A are not B*.
The model proposes three main heuristics. The min-heuristic results in choosing the quantifier from the least informative premise (the min premise) for the conclusion (which is similar to the atmosphere effect). This heuristic "entails" some other conclusions with a certain probability which is called p-entailment. This p-entailment is the second applied heuristic, e.g. ,*All A are B* leads to the probable enclosed conclusion that also *Some A are B*. A third heuristic, the attachment-heuristic, determines the ordering of the selected quantifier. The subject of the min premise is used as the subject of the conclusion if it

is an end term (i.e. if it does not occur in both premises), otherwise the end-term of the max premise is used.
The generated conclusions are tested with two additional heuristics. The max-heuristic states that the probability of the correctness of a conclusion correlates with its informativeness. The O-heuristic goes one step further and states to avoid conclusions of the O-type.
The application of well defined heuristics and the prediction in confidence for different syllogisms are one of this theory's advantages. These heuristics however fail in explaining an increase in logical correctness and they do not account for the vast individual differences (Khemlani & Johnson-Laird, 2012). One might therefore argue that this theory is not generally applicable but rather to a set of individuals who rely on probabilistic reasoning strategies.

2.2.6 The mReasoner

The mReasoner is a computational theory implemented in the programming language *Lisp*. The implementation is based on building mental models of syllogistic premises and the use of four different parameters which are explained in more detail below (Khemlani & Johnson-Laird, 2013).

1. λ: This parameter denotes to the size of the model, i.e. it represents the number how many entities are represented in the mental model which is drawn from a Poisson distribution with parameter λ. Hence a bigger λ increases the probability of a higher number and a greater variety of such.

2. ϵ: The ϵ parameter (canonicality) sets the probability to which a reasoner considers the whole set of possibilities in their mental model. E.g. *All A are B* contains A's that are all B's but also B's that aren't A's. A low setting of ϵ omits the latter set in the mental model.

3. σ: With this parameter, one can set the possibility of the model to search for counterexamples. This increase the *deliberative* part of the model.

4. ω: The last parameter ω is important when a counterexample is found, so it is dependent on σ. It sets the model's probability of weakening an initial drawn conclusion when a counterexample is found. E.g. when initially the conclusion *All A are C* has been drawn, weakening the conclusion would lead to *Some A are C*.

Comparing the generated data against empirical, Khemlani and Johnson-Laird established three different settings for the mReasoner to cluster three types of reasoners. They could distinguish between deliberative reasoners who make use of counterexamples, intuitive reasoners rarely using counterexamples and also a mixture of both types, using canonicality und counterexamples to some degree. The same clustering could also be applied by Ragni et al. (2018) on a set of data and yielded decent results.

2.3 Learning

When examining effects of learning it is necessary to get a clear understanding of what learning is and how it can be described. The word learning can be seen as "The acquisition of knowledge or skills through study, experience, or being taught." (Simpson, Weiner, & Oxford University Press, 1989) which resembles the understanding of learning in everyday language. Psychologically there are multiple understandings of the term learning. One can distinguish between primary and secondary cognitive skills which can both be acquired. Primary cognitive skills are denoting to cognitive abilities that evolve from biological development whereas secondary cognitive skills are learned via culture e.g. in schools (Sweller, 2008).
Considering Piaget's work on learning in infants reasoning can be classified as a primary cognitive skill which however involves learning and development of the brain. This process of reasoning demands the ability to create abstract representations and mental images of the world and incoming information in order to draw inferences. At about the age of 10 years, children are able to draw simple conclusions of the form: *Max is older than Sally and Sally is younger than Daniel. Which one of them is the oldest?* (Piaget, 1936).
When looking at the process of learning itself it can be said that learning arises from the demands of the environment. Those demands make a development possible which leads to a permanent transformation in either behavioral aspects, knowledge or personality. This transformation can only occur when the individual is dealing actively with the information it is confronted with and is interpreting the incoming sensory stimuli in a way that it connects with prior experiences. The individual is thus constructing its own schemata which is the most important part of the learning process (Anderson, 2013; Nückles & Wittwer, 2014). With creating these schemata and integrating new information with prior knowledge from the long-term memory it is possible to build bigger chunks of information which ultimately leads to a relief of the working memory allowing it to process data faster

or easier. On one hand learning can therefore be understood as constructing concepts from knowledge which is why this theory of learning is called constructivism. In constructivism knowledge is seen as an abstract entity and learning takes place in three processes, namely:
1. selection of information which can be directed consciously
2. organization of the incoming information
3. integration of the information with prior knowledge of the long-term memory

On the other hand there is the theory of situated learning. In this case learning can be seen as becoming part of a community. Knowledge is therefore bound to the situation at hand and learning takes place via social exchange (Nückles Wittwer, 2014).

In the field of computer science one can often see the distinction between supervised and unsupervised learning. In this case supervised learning refers to the presence of feedback. Unsupervised learning on the contrary means that there is no explicitly labeled feedback, e.g. learning to recognize "good" and "bad" traffic days without ever getting told which day was good or bad (Russell, 2015). A similar distinction can also be made in psychology where in supervised learning "the learner is given a stimulus, classifies it, and is provided with corrective feedback" (Love, 2002) and in unsupervised learning there is no such feedback. Unsupervised learning usually leads to a stronger categorization of concepts and objects and the more prior knowledge there is the stronger this phenomenon occurs (Kaplan Murphy, 1999).

In any case, learning seems to follow a predictable course, called the *power law of learning*. It states that when acquiring new skills the increase follows a logarithmic distribution. I.e. that in the beginning the improvement is big and then flattens over the course of time converging to a maximum. This phenomenon seems to be independent of the examined domain, either in perceptual-motor skills (Anderson, 2013) or in cognitive abilities such as memory or abstract problems like mental rotation of objects. (Heathcote, Brown, & Mewhort, 2000). Figure 3 illustrates the power law of learning.

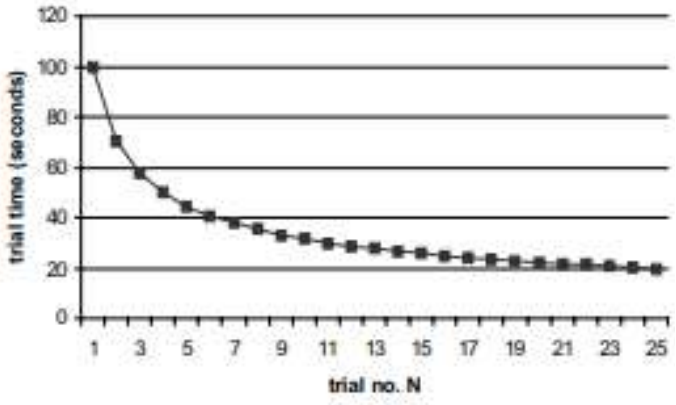

Figure 3: Changes in reaction times that follow the power law of learning (Roessingh, Hilburn, Nationaal, & Ruimtevaartlaboratorium, 2000).

2.3.1 Learning in Reasoning

As stated earlier human reasoning can take on several forms. Although it is still debatable if there is a clear differentiation between a type I approach using heuristical processes and a type II approach relying on analytical thinking it is certainly true that heuristics are in their nature more prone to errors (since this is among others what characterizes heuristics). And if the goal is to learn reasoning (i.e. in this case to get better and more correct) it seems to be an obvious approach, despite being able to tell whether humans are using heuristics more often or not, to engage the learner in using lesser heuristics and therefore type I processes and instead use more of analytical type II processes (Le Wartschinski, 2018). By evoking certain reasoning schemata encouraging these processes Cheng Holyoak (1985) could show that participants were able to solve reasoning tasks more correctly. Regarding syllogistic reasoning in particular, there have been approaches examining learning within this domain. It could be shown that human reasoning can be improved with a learning program containing different methods, e.g. Venn Diagrams in younger children. After completing these learning programs, they scored significantly higher in syllogistic reasoning tests (Lane, Fletcher, Fletcher, 1983). Another study from Johnson-Laird and Steedman (1976) tested participants on two separate occasions approximately a week apart with all 64 possible combinations of premises. The participants were asked to draw their own conclusions from each of the premises. A significant increase in correct answers was found from week 1 to week 2 revealing a positive learning effect.

The study could also show that the time which participants needed correlated negatively with giving the right answer, meaning that the lesser time the subjects needed the answer was more likely to be correct. What is noteworthy is that the subjects did not know they were being tested a second time beforehand. It can thus be concluded that this learning effect results from experience the participants made while doing the tasks for the first time.
The study could also show that reasoning in itself is not a stable process, intra-individually as well as inter-individually. Participants changed their answers from 15 up to 39 different answers regarding week 1 to week 2 which in regard to the increase in correct answers can also be taken as an indication for a learning process (Ragni et al., 2018). Bucciarelli and Johnson-Laird (1999) also examined the performance on syllogistic reasoning tasks on two sessions with a week apart from each other. In this case the participant were instructed to use mental models for the tasks either in the first or in the second week for a counter-balanced design. Opposing to the results of Johnson-Laird and Steedman the accuracy in terms of correctness did not increase for both groups. The use of mental models did however lead to a significantly more diverse set of conclusions.
One might therefore suggest that imposing a strategy to use for reasoning tasks might hinder the learning process. In this experiment participants also showed a large variety in the actual use of the mental models which might suggest that there is a certain bias towards sticking to own reasoning strategies.

2.4 The Current Study

Logic and reasoning in humans have been topics of research for a long time and syllogistic reasoning in particular has been a means for analyzing this field of research in a great amount of studies. The goal of the current study is to gain new insights in syllogistic reasoning and especially in learning effects in this domain. Although Johnson-Laird and Steedman (1978) tested participants in syllogistic reasoning twice and could show that there have been differences no study has tried to extend this approach and hypothesizing that there might be yet other differences occurring. Therefore in this study it will be examined how participants differ in their response behavior when being tested in syllogistic reasoning for a total of three times and especially if there are significant differences from the second to the third testing.

Taking the previous research and general insights of learning into account, the following research questions will be examined:

1. Following the results of Johnson-Laird and Steedman's study participants should be giving significantly more correct answers in the second testing and improve even more in the third testing.

2. Learning effects should result into the creation of more and stable schemata which leads to a faster processing of information. Therefore the response latency should decline from testing to testing which would also be in accordance with the findings of Johnson-Laird and Steedman's study.

3. Another effect of creating schemata should be that the reasoning process in the person itself becomes more consistent. This means that participants should be making significantly less changes in their responses from the second to the third testing than they did from the first to the second testing.

    3.1. Assuming that the selected theories are describing a strategy which people use to solve syllogistic tasks, a theory that predicts one person's performance in the first testing well should also do so in the following testing and presumably be able to predict the person even better then.

Chapter 3

Method

3.1 Design and Sample

To answer the underlying hypotheses an experiment with a total of 28 participants was conducted. To observe the individual changes of participants in syllogistic reasoning, the experiment's design has been a within-subject design with one factor, namely the number of testings and hence the elapsed time from testing to testing.
The sample consisted of 28 students, recruited directly in lectures and via flyers, who were either freshmen studying cognitive science as their minor subject with mixed main subjects or other students whose study programs focused mainly on psychological issues. People with prior knowledge to syllogisms were excluded from the experiment. The subjects participated either in order to fulfill their requirements for their study program or have been compensated with 15 € for their participation. The participants were between 18 and 29 years old (M = 21.6 years, SD = 2.6), 9 of them were male and 19 female.

3.2 Material and Procedure

Johnson-Laird and Steedmann (1978) made some suggestions for experiments with syllogisms. E.g., a syllogistic experiment should contain a sufficient amount of problems. So for this experiment all possible combinations of moods and figures have been realized in the experiment which made up a total of 64 syllogisms. Another prerequisite is the context independence of statements. Each premise should be presented in a meaningful but as neutral as possible linguistic context since there have been studies suggesting that an abstract representation of problems leads to a distortion of performance. Hence in this study neutral affiliations like 'diver' or 'baker' have been chosen for the subjects of the premises. The possible conclusions which the participants were presented and from which they could choose, have also been constructed from all the possible combinations of moods and figures. Since the conclusion of a syllogism can either make a statement

about the relation of a to c or the other way round, the following options of answers are possible:

$$Figures = \{ac, ca\}$$
$$Moods = \{A, E, I, O\}$$
$$\|Figures\| * \|Moods\| = 8$$

And additionally, the answer option *no conclusion is possible* was contained which led to a total number of nine answer options. An example task therefore looked like this:

Premise 1: No surfers are engineers
Premise 2: Some engineers are artists

Figure 4: Experimental setup for an example task

The cognitive reflection tasks have been taken from *Cognitive Reflection and Decision Making* (Frederick, 2005) so that they reliably predict general cognitive ability. These tasks consist of assertions that depict a scenario where simple mathematical properties are involved and are designed in such a way that a prepotent response that is mostly incorrect occurs at first and should be overridden in order to find the correct response. One item for example has been:

"A bat and a ball cost $1.10 in total. The bat costs $1 more than the ball. How much does the ball cost?"

People tend to give the answer 10 cent at first and have to reflect the given answer to realize that this is wrong (Toplak, West, Stanovich, 2011).
The subjects were tested alone or in groups and had been told that they were participating in an experiment researching logic and its development in humans that are solving logic

tasks over a longer period of time. That means they were informed that there would be a second and a third testing. The experiment was done on customary laptops and the participants had been instructed to answer quickly and as correct as possible. Their task was to choose a conclusion that necessarily follows from given syllogistic premises. In order to get used to the format an example with only one premise was presented before the real testing of the 64 syllogisms began. The syllogisms were presented in a random order with the two premises always shown on the screen as well as the nine possible answer options that were all visible at once and selectable via the corresponding key on the keyboard.
In the first testing seven different cognitive reflection tasks were presented after the 64 syllogisms to assess general cognitive ability. Approximately one week after the first testing the participants were tested for a second time and again for a third time after another two weeks. The experiment in the second and the third testing were identical to the first one and contained the same syllogisms but the cognitive reflection tasks were cut out. Each session lasted about 45 to 60 minutes on average.

Chapter 4

Results

For analyzing the subjects' performance all of their answers to the 64 syllogisms were considered.

4.1 Correct Answers

One of the assertions was that the subjects improve their performance on the reasoning tasks which means that they should get more correct answers from testing one to testing two and from the second one to the third one.
To test whether the participants improved, the mean number of correct responses were compared. Figure 5 illustrates the mean number of correct responses for every testing as an overall score as well as separated into valid syllogisms, i.e. syllogisms which can be drawn a logically valid conclusion from and into invalid syllogisms where no valid conclusion is possible.

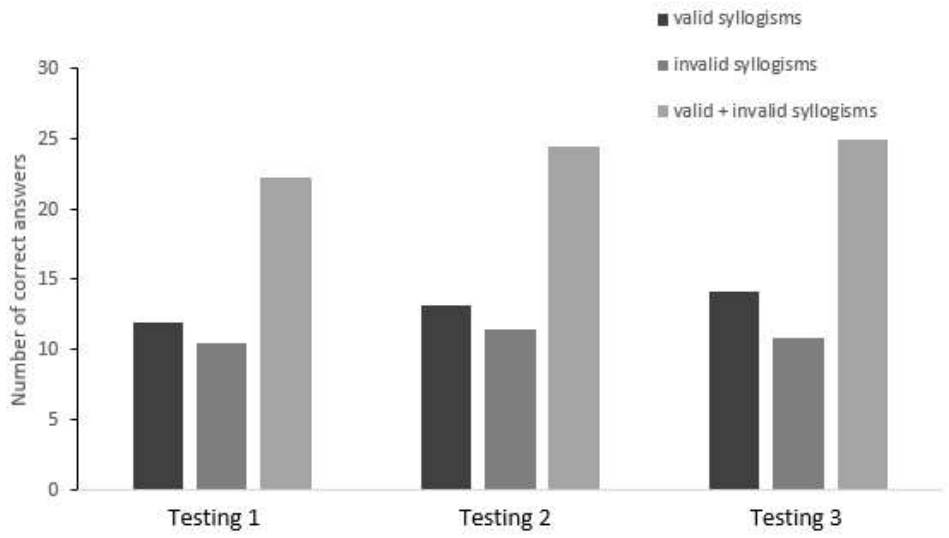

Figure 5: Mean number of correct answers for each testing. Note that from the 64 total syllogisms there are 27 valid and 37 invalid ones

For comparing the results the Page's trend test was used because it is explicitly useful to measure monotonous orderings, especially when there are three groups or more. Such effects where an ordering among the treatment groups is expected, are typically neglected when comparing the results with variance analysis (Page, 1963).

The Page's trend test for the number of correct responses for all syllogisms yielded no significant differences between the testings, L = 346, $p > 0.05$. A significant difference was however found for the number of correct answers on valid syllogisms, L = 357, p = .003. That means that subjects increased their performance on valid syllogisms and answered more of them correctly from testing to testing whereas there was no performance increase to be found for invalid syllogisms.

Another remarkable point is the huge inter-individual difference between the subjects. The number of correctly given answers ranged from 3 correct answers to 59 correct answers, of which the latter was found in the third testing.

Big differences were also found between the syllogisms. E.g., the AE2 syllogism was answered correctly by a single person in the first testing whereas other syllogisms, e.g. AI2 seemed to be feasible for most participants (22 correct answers).

As prior research suggests, there should also be differences between the figures of syllogisms. Hence, the tasks were divided into the different figures which can be seen in Table 1. A Friedman ANOVA yielded a significant difference between the correct answers for the different figures in the third week ($\chi^2(3) = 15.738$, p = .001). A post-hoc test using Bonferroni correction revealed that the participants were significantly more correct on syllogisms of Figure 1 than they were on syllogisms of Figure 3 and 4.

	Number of correct responses		
	Testing 1	Testing 2	Testing 3
Figure 1	5.3	5.5	5.5
Figure 2	5.1	5.8	5.6
Figure 3	6.0	6.8	6.7
Figure 4	5.9	6.4	7.1
Total	22.3	24.5	24.9

Table 1: Mean number of correct answers subdivided into the different figures. A significant difference was found for the third testing between Figure 1 and Figure 3 (observed rank difference = 27.5, critical difference = 25.48) and between Figure 1 and Figure 4 (observed rank difference = 29.0, critical difference = 25.48)

Furthermore, as expected there was a high correlation between the correct answers on the cognitive reflection tasks and the correct answered syllogisms (Spearman correlation, r = 0.64, p < .001).

4.2 Reaction Time

The time needed to work on the syllogistic tasks was analyzed by observing the reaction times for each syllogism. Page's trend test also yielded a significant difference for the mean of the reaction times (L = 372.0, $p < .001$) meaning that the participants have become progressively faster in each testing.

4.3 Consistency

The third hypothesis was that subjects will perform more consistent over time. In order to analyze the consistency of answers, the number of differing answers from the first to the second and from the second to the third testing was regarded which will be called shifts. In Figure 6 the total number of shifts for each participant can be seen.

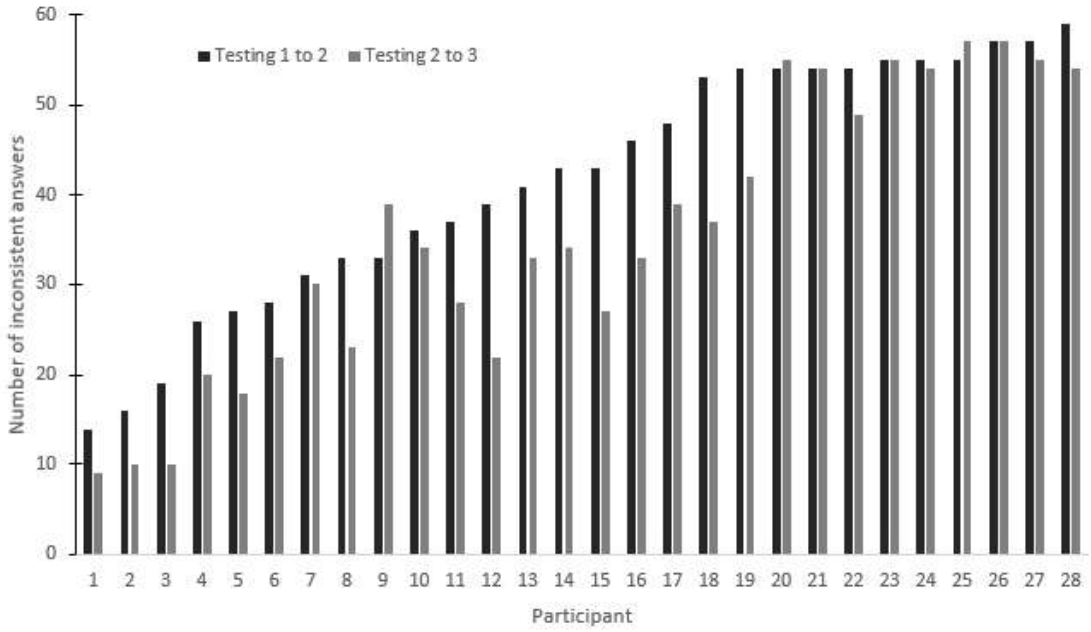

Figure 6: Number of inconsistent answers for each participant from the first to the second testing and from the second to the third testing

A Wilcoxon rank test for the total number of shifts yielded a significant result, Z = 18.5, $p < .001$, which means that the participants changed significantly less answers from the second to the third testing than they did from the first to the second testing.
Looking at the syllogisms divided into the different Figures differences between them became apparent.

	Number of shifts	
	Testing 1 to 2	Testing 2 to 3
Figure 1	9.3	7.9
Figure 2	10.4	8.8
Figure 3	11.3	10.2
Figure 4	10.8	8.8
Total	41.8	35.7

Table 2: Mean number of inconsistent answers from the first to the second testing and from the second to the third testing for each figure

Friedman's ANOVA showed that there was a significant difference between the figures for the shifts from testing one to testing two ($\chi^2(3) = 8.37$, p = .039) and also in the shifts from the second to the third testing ($\chi^2(3) = 14.64$, p = .002). In both cases the significant difference was found between Figure 1 and Figure 3 which was calculated using post hoc tests with applied Bonferroni corrections. The critical difference in both cases was 25.48 and the observed rank difference between Figure 1 and Figure 3 were 26.5, respectively 34.
Once more, big individual differences can be observed between the subjects. As seen in Figure 6 individuals changed up to 57 of their answers from the second to the third testing meaning that almost every answer was changed whereas 9 was the lowest number of shifts which means that almost every answer was consistent with the one given in the previous testing.
Lastly it was assumed that the subjects would get more engaged in one problem solving strategy. Therefore the performance of the subjects was matched with the prior described theories of syllogistic reasoning, namely the atmosphere theory, the PHM theory and the mental models theory. Each of these theories predict a set of answers for each of the 64 syllogisms. Since the theories predict an inconsistent quantity of answers, e.g. the atmosphere theory predicts two different answers for the AA1 syllogism whereas the PHM theory predicts a set of four possible answers for the same syllogism, comparing them was not possible without adjustment. This is because the best theory in this case would be

one that predicts every possible answer. Hence the precision measure was used to achieve comparable results using the following formula.

$$Prediction\ rate = \frac{\sum_{i=1}^{n} hits_{syllog_i}}{s}$$

With s = number of total predictions for one theory and n = total number of syllogisms. This means the number of correctly predicted answers is divided by the total number of answers predicted by the corresponding measure.
To show how well the theories performed against each other, the amount of how often one theory predicted the answers for one participant more accurately than the other theories is illustrated in Figure 7.

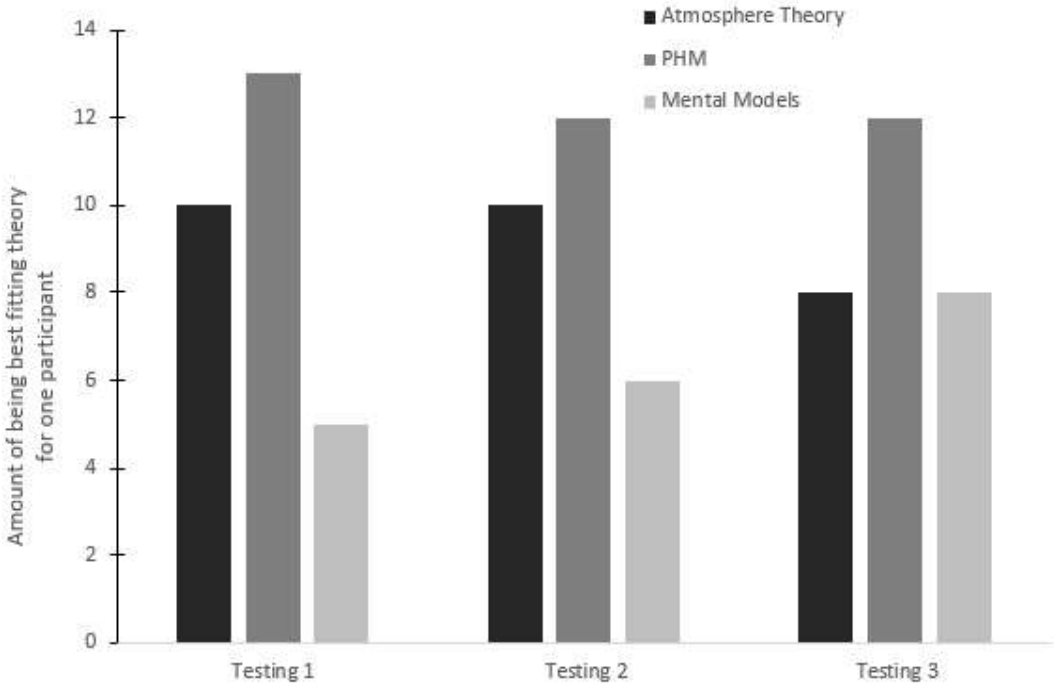

Figure 7: Number of times each theory predicted the responses best

As can be seen in Figure 7, the most frequent best theory has been the probability heuristics model (PHM) and the atmosphere theory has been the least predictive theory. The mental model theory gained an increase in prediction from seven best predictions in week one to nine in week three. No significant increases could be found for any theories' prediction rates ($p > .05$).
Furthermore, it could be observed that most of the participants (n = 18) switched between

theories and did not show a consistent engagement in one theory. For a better illustration the theories' prediction rates are depicted in a heatmap (s. Figure 8) with a color coding containing every theory and every participant. A darker color means a better prediction rate.

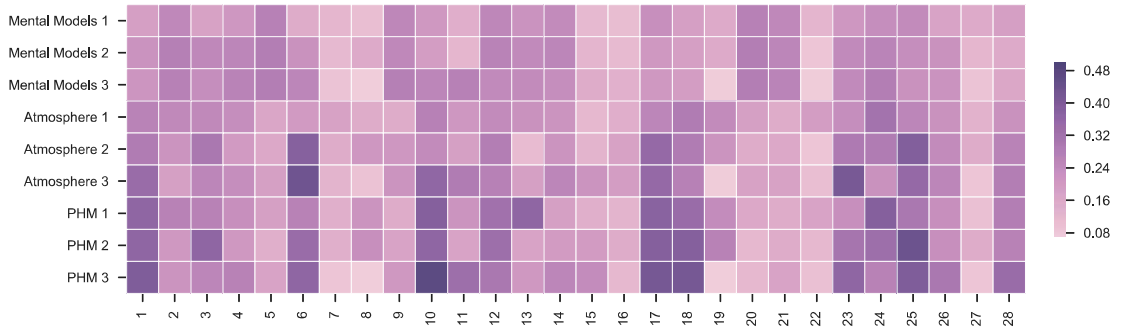

Figure 8: Heatmap for the prediction rates of Atmosphere, PHM and Mental Model theory. The number next to the theory's name stands for the number of the testing

The heatmap shows that the PHM theory had high prediction rates in many cases, which is also true for the Atmosphere Theory. The Mental Models theory doesn't have these high prediction rates but also has fewer prediction rates that are very low.
Supporting this, in Figure 9 the maximum prediction rates are illustrated for every theory for every testing. That means the highest prediction rates across all subjects was taken for every theory. It can be seen that the Mental Models Theory gained only little increase in comparison to the other theories.

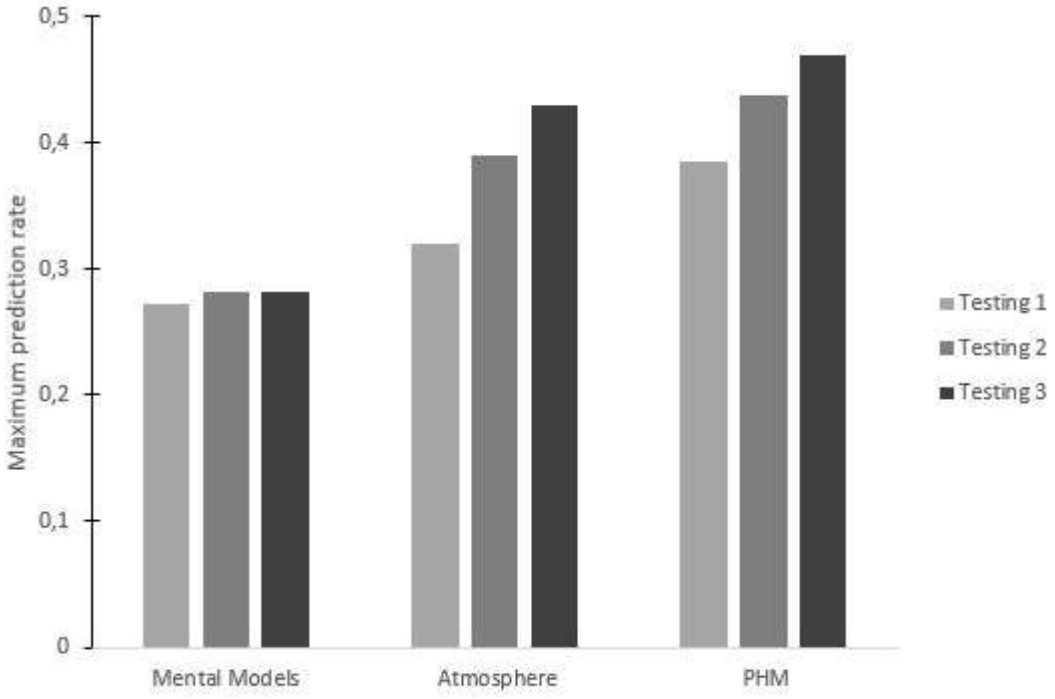

Figure 9: Overall maximum prediction rates for every theory

The experimental data was furthermore analyzed and compared with synthetically created data from the mReasoner. Therefore three different settings were applied, which can be seen in Table 3. For each setting, 28 datasets were created simulating the answers for all the 64 syllogisms. Each person was then assigned one of the datasets randomly.

Settings of the mReasoner parameters				
Clusters	λ: Size of model	ε: canonicality	σ: search for counterexamples	ω: weaken conclusion
Intuitive	2.0	0.0	0.4	0.6
Deliberative	3.0	0.6	0.8	0.6
Mixture	2.0	0.0	1.0	0.8

Table 3: Settings for the mReasoner

Figure 10 shows a heatmap illustrating the prediction rates of the different mReasoner settings. The heatmap shows that some participants are described very well by deliberative or mixed strategies and that high prediction rates in the first testing continue for the other testings. Furthermore, there are many participants that show very low prediction

rates for each setting.

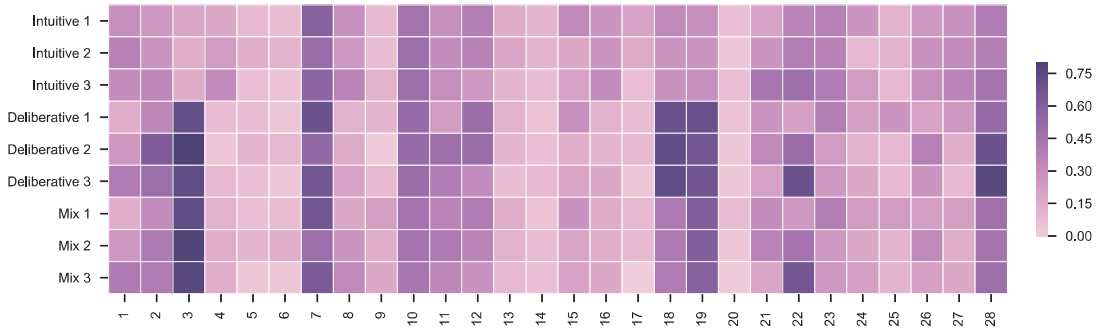

Figure 10: Heatmap for the prediction rates of the three different settings for the generated mReasoner data

For the mReasoner the best predicting settings were also analyzed just like for the different therories before. The intuitive setting was the best predictor for the first and the third testing whereas the mixed setting was the least predictive setting for all the testings.

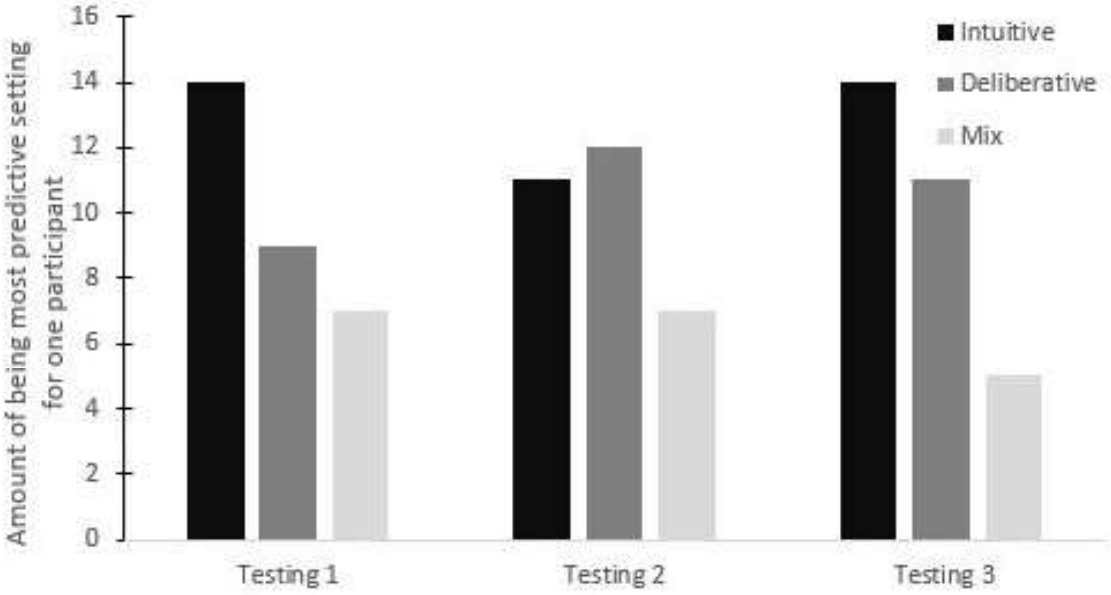

Figure 11: Number of times one setting of the mReasoner predicted the responses best

Chapter 5

Discussion

Two big questions have been examined in the course of the conducted study: How do people reason and how does the reasoning process change over time? Several insights can be gained from the experiment.

First, in contrast to the experiment of Johnson-Laird and Steedman (1978), participants did not increase their overall correctness on the syllogistic reasoning tasks. Yet, after dividing the tasks into valid syllogisms which a conclusion can be drawn from and into invalid syllogisms where logically no conclusion follows from the premises, an improvement was observable.

Interestingly, the performance increase in Johnson-Laird and Steedman's study was mainly due to the increase on valid syllogisms as well and a decline could be observed for the tendency to answer "no valid conclusion" on syllogisms of figure 4. A similar finding is yielded by the current study where a constant increase and the biggest total increase of 1.2 more correct answers in average could be found for figure 4. This could reflect the fact that the improvement was not as large as in the earlier study - a fact that may be a result of the different samples of participants and the slightly different experimental designs. While in Johnson-Laird and Steedman's study participants were instructed to draw conclusions on their own, participants in this study had to select an answer from the whole set of possible conclusions. Studies suggest that choosing from a set of given answers might lead to a more surficial learning approach and vice versa, generating own content can lead to the employment of deeper learning strategies (Scouller, 1998). But nonetheless, a constant improvement could be observed which shows that the learning process is ongoing and is probably converging to a maximum in a logarithmic curve, accordingly to the power law of learning.

Furthermore it is apparent that valid syllogisms were more feasible from the beginning for the participants than invalid ones. Despite there being 10 more invalid syllogisms in total, less of them were answered correctly. So the improvement for valid syllogisms could be accounted for by the stabilization of the prior handled tasks. Since these tasks might have been easier for most participants, upon reflection afterwards the easier tasks could

have led to a stabilization and a bigger confidence for them, leading either to the belief that there has to be a conclusion for all syllogisms or an interference since these ground foundations exert a stronger emergence. It remains however open if participants were correct in their assertions in the first place. An earlier study examining metacognition could show that people tend to be overconfident and that their self-evaluation is rather imprecise (Bajšanski, Močibob, & Valerjev, 2014), however this study did not examine all syllogisms and can therefore not exhaustively answer the question whether some syllogisms lead to a higher confidence.

Besides this, the experimental situation in itself can lead to a non-neglectable change in the behavior of participants in a way that they feel compelled to find a conclusion to every task and they don't want to end up finding no one.

The number of correct answers also hints at big individual differences which is consistent with previous findings (Khemlani Johnson-Laird, 2012). There are people who tend to experience more trouble with syllogistic reasoning whereas other people scored almost perfectly.

Since reflection on cognitive processes, as tested by the cognitive reflection tasks, are an indicator for general cognitive ability and intelligence, the high correlation between these tasks show that syllogistic reasoning demand cognitive ability. CRTs also show a high prediction rate for heuristic and biased tasks and can depict the propensity to miserly processing (Toplak et al., 2011). The performance on CRTs could thus also hint at a general need for cognition, i.e. the joy of solving the given task and the willingness to use cognitive resources which in turn correlates with intelligence (A. R. Cohen, Stotland, & Wolfe, 1955).

The fact that subjects got progressively faster supports the theoretical claim of creating schemata and connecting new information with prior information. It could be argued that the participants were less willing to invest cognitive and time resources with time but since an improvement could be observed this doesn't hold up and this finding can clearly be interpreted as a learning effect.

Usually in motoric skills learning often involves an automation of processes (Anderson, 2013).An automation process probably occurs in a similar way when reasoning is learned. Considering the different theories, respectively reasoning strategies described earlier, the learning process could result in the automation of heuristics, the search for counterexamples or the construction of mental models. Let us consider the latter and hence the mental model theory. Through learning the semantic interpretation might be sped up or the semantic interpretation and combination of the two premises could be condensed into one step, similar to *production compilation* which is used to model experience gain

in cognitive modeling. (Taatgen & Lee, 2003).

5.1 Theories of syllogistic reasoning

One main problem of comparing the theories is the different amount of answers they predict. When only validating whether a theory contains a given answer in its set of predictions, the best theory would be the one that predicts everything. That given the case the applied metric created comparable outcomes.
The fact that the probability heuristics model was the best predicting theory in all testings reflects the fact of the general low amount of correct answers (34.8%, 38.3%, 38.9%) compared to the study of Johnson-Laird and Steedman (1978) (58%, 68%) and also the use of heuristic reasoning strategies. The decrease in predictive power and the concomitant of a better prediction of the mental model theory could indicate an increased application of logical rules. This insight should be treated with caution though since the differences were subtle most of the time.
A more striking observation is that most participants answers could not be predicted confidently with one theory and shifting between theories happened which was not in accordance with the hypothesis that people use one strategy and get more engaged into it. Considering the heat map several clusters can be grouped: The expected behavior can be seen in some cases and the prediction rates got better every testing for one theory. For other participants no theory could provide a satisfactory prediction and a third cluster could contain participants that shifted between theories.
Similar results were yielded by the analysis of the mReasoner program. In the first testing the intuitive settings could predict most of the participants answers which is in line with the fact that the PHM theory was the best predicting theory for testing number one. Although not constant an increase for the deliberative setting was found which also indicates the use of logical rules, like finding counterexamples. It should be noted that the differences in prediction rates are again subtle and should therefore be treated cautiously. The heatmap for the mReasoner settings also illustrates clusters where some participants are described very accurately, others that cannot be described well and the third cluster of participants where a shift between the settings seemed to occur.
Regarding the predictions of the mReasoner and the different theories, different types of

reasoners are proposed that are not dependent of the theory that describes their reasoning best but rather their development in reasoning tasks:

1. Strong reasoners: These reasoners are confident in their reasoning abilities following a certain strategy that is to some degree describable through one theory.
2. Insecure reasoners: These reasoners may not be as confident in their judgments or in their problem-solving approach and upon reflecting they may be deciding to use another strategy.
3. Weak reasoners: This group may be the most difficult one to describe - and apparently the most debatable one. This group is not well describable through the theories which may be due to using different methods for different syllogisms - which should however change after several testings or due to other reasons, e.g. giving up in trying to find an answer.

One of the most important findings of this study was the significant decrease in the insecurity of given answers. Since no study so far has examined syllogistic reasoning over such a long period of time, reports of answer behavior were restricted to the comparison of one testing to the other. In Johnson-Laird and Steedman's study (1978) an instability in the reasoning process could be determined through analyzing the number of inconsistent answers. Comparing this to the shifts of answers of the current study a similar tendency can be observed for the shifts from the first to the second testing which can be seen in Figure 12.

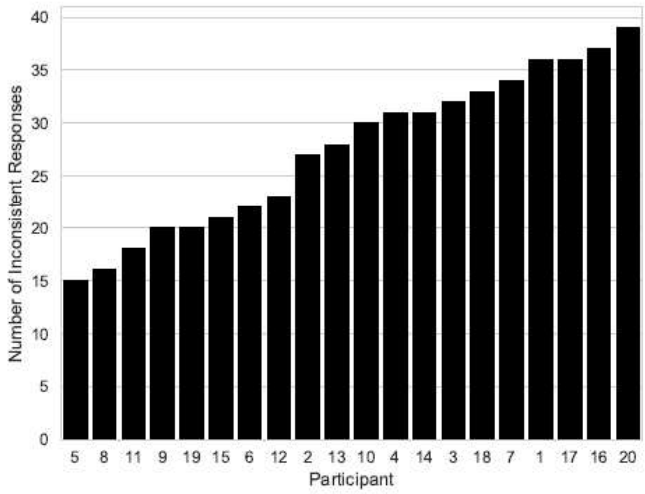

Figure 12: Number of inconsistent answers for the experiment of Johnson-Laird and Steedman (Ragni et al., 2018)

These shifts had a big range in both studies and when ordered a continuous increase becomes clear. This is once more evidence for big differences between individuals and more so evidence that reasoning even within people is unstable. This study however could show a significant decrease in instability from the second to the third testing. Only three of the participants changed more answers from the second to the third testing than they did from the first to the second one.

Considering syllogistic research this is arguably an important finding. Most research and theories aim to describe the process of reasoning universally. However a big question arises with the findings of this study. Is the predictive power of any theory of reasoning justifiable when answers from one person seem to underlay instability in such a manner and need time to get stable?

The results suggest that the there is high fluctuation and a lot of noise in the system of reasoning. Even the most stable subject of this experiment changed 14 of their answers in the first testing. The reduction of this number indicates that outcomes of reasoning processes are stabilizing with repeated exposure.

Hence, it is proposed that experiments examining reasoning or at least syllogistic reasoning in particular should take this finding into account. To reduce the noise in the system participants should be tested over several occasions. This is of course an idealized proposal since resources are limited and people are not available for as long as one wants them to be. But a sufficient amount of testings should be considered to get reliable results at least.

Thus when proposing hypothetical thoughts to describe syllogistic reasoning, a good theory should therefore be able to account for development of people's reasoning and be able to do so for different types of reasoners - something none of the theories can provide so far. A subtle but impactful adjustment of the current theories would be to give a specific weight to every predicted answer. This would lead to

(a) The possibility to predict changes in answering behaviour through adjusting the weight of the answers. E.g. some syllogisms demand the search for counterexamples (Ragni et al. 2018). Learning outcomes could therefore be modeled by giving more weight to conclusions that are yielded by finding counterexamples - assuming that a learner strengthens the ability to find counterexamples over time.

(b) Applying a new metric on the weighted scores would make it comparable to other theories such as mReasoner.

Let us consider the precision measurement used in information retrieval. The precision at spot k of a result list is defined as the number of hits until k divided by k. (Manning, Raghavan, & Schütze 2008)
Let the results list be equal to the list of predictions a theory makes. Let k be the position of a hit, i.e. the position at which the given answer coincides with the predicted answer and s be the number of predictions for a given syllogism. Then the following formula for the precision of a syllogism could be used to account for the weighting from which an average precision for n syllogisms can be calculated:

$$Precision_{syllog} = \frac{Precision@k}{s}$$

$$Average\ Precision = \frac{\sum_{i=1}^{n} Precision_{syllog_i}}{n}$$

If the answer is not in the predicted set, let $Precision@k = 0$

It can be noted that modeling people's reasoning strategies is difficult and predicting their development might just be as hard. If there are different types of strategies used, it is only natural that learning occurs differently which is a challenging task to model. The mReasoner in this instance is ahead of other theories being able to model changes through adjusting the different parameters but is still lacking a decisive explanation for learning effects.

Chapter 6

Conclusion

Are people learning reasoning over time and do the get better at solving logic tasks? They probably do and they surely show changes in their answering behaviour.
Is there one single theory capable of describing how every human reasons? Definitely not. The presented study indicates that reasoning is unstable and that there is a large variety in the reasoning process itself between individuals and also within their own development when they face several reasoning tasks over a longer amount of time.
Theories describing syllogistic reasoning should be able to account for the observed phenomena - a property which almost every theory lacks at the moment.
It is therefore highly suggested for future research, that examines syllogistic reasoning, to test individuals over several occasions. Future studies have yet to find a satisfactory answer for the development of reasoning - and a theory that claims to have decrypted the unknown processes of reasoning has to be able to accommodate its predictions to a reasoner and their individual development.

References

Aerts, D., Gabora, L., Sozzo, S., & Veloz, T. (2011). Quantum Structure in Cognition: Fundamentals and Applications. ArXiv:1104.3344 [Quant-Ph]. Retrieved from http://arxiv.org/abs/1104.3344

Anderson, J. R. (2013). Cognitive Skills and Their Acquisition. Psychology Press.

Bajšanski, I., Močibob, M., & Valerjev, P. (2014). Metacognitive Judgments and Syllogistic Reasoning. Psihologijske Teme, 23(1), 143–166.

Bucciarelli, M., & N. Johnson-Laird, P. (1999). Strategies in Syllogistic Reasoning. Cognitive Science, 23, 247–303. https://doi.org/10.1207/s15516709cog23003_1

Chater, N., & Oaksford, M. (1999). The probability heuristics model of syllogistic reasoning.
Cognitive Psychology, 38(2), 191–258. https://doi.org/10.1006/cogp.1998.0696

Cheng, P. W., & Holyoak, K. J. (1985). Pragmatic reasoning schemas. Cognitive Psychology, 17(4), 391–416. https://doi.org/10.1016/0010-0285(85)90014-3

Cohen, A. R., Stotland, E., & Wolfe, D. M. (1955). An experimental investigation of need for cognition. The Journal of Abnormal and Social Psychology, 51(2), 291–294. https://doi.org/10.1037/h0042761

Cohen, J. (1994). The earth is round (p<.05). American Psychologist, 49(12), 997–1003. https://doi.org/10.1037/0003-066X.49.12.997

Cosmides, L. (1989). The logic of social exchange: Has natural selection shaped how humans reason? Studies with the Wason selection task. Cognition, 31(3), 187–276. https://doi.org/10.1016/0010-0277(89)90023-1

Evans, J. S. B. T., Newstead, S. E., & Byrne, R. M. J. (1993). Human Reasoning: The Psychology of Deduction. Psychology Press.

Frederick, S. (2005). Cognitive Reflection and Decision Making. Journal of Economic Perspectives, 19(4), 25–42. https://doi.org/10.1257/089533005775196732

Grice, H.P. (1978). "Further Notes on Logic and Conversation," Syntax and Semantics, vol.9 edited by P. Cole, Academic Press. Reprinted as ch.3 of Grice 1989, 41–57.

Griggs, R. A., & Cox, J. R. (1982). The elusive thematic-materials effect in Wason's selection task. British Journal of Psychology, 73(3), 407–420. https://doi.org/10.1111/j.2044-8295.1982.tb01823.x

Heathcote, A., Brown, S., & Mewhort, D. J. K. (2000). The power law repealed: The case for an exponential law of practice. Psychonomic Bulletin & Review, 7(2), 185–207. https://doi.org/10.3758/BF03212979

Herrmann, E., Call, J., Hernàndez-Lloreda, M. V., Hare, B., & Tomasello, M. (2007). Humans Have Evolved Specialized Skills of Social Cognition: The Cultural Intelligence Hypothesis. Science, 317(5843), 1360–1366. https://doi.org/10.1126/science.1146282

Johnson-Laird, P. N. (1978). The Meaning of Modality. Cognitive Science, 2(1), 17–26. https://doi.org/10.1207/s15516709cog02012

Johnson-Laird, P. N., Oakhill, J., & Bull, D. (1986). Children's Syllogistic Reasoning, Children's Syllogistic Reasoning. The Quarterly Journal of Experimental Psychology Section A, 38(1), 35–58. https://doi.org/10.1080/14640748608401584

Johnson-Laird, P.N., & Steedman, M. (1978). The Psychology of Syllogisms. Cognitive Psychology - COG PSYCHOL, 10, 64–99. https://doi.org/10.1016/0010-0285(78)90019-1

Johnson-Laird, P. N. (2010). Mental models and human reasoning. Proceedings of the National Academy of Sciences, 107(43), 18243–18250. https://doi.org/10.1073/pnas.1012933107

Kaplan, A. S., & Murphy, G. L. (1999). The acquisition of category structure in unsupervised learning. Memory & Cognition, 27(4), 699–712.

Khemlani, S., & Johnson-Laird, P. N. (2012). Theories of the syllogism: A meta-analysis. Psychological Bulletin, 138(3), 427–457. https://doi.org/10.1037/a0026841

Khemlani, S., & Johnson-Laird, P. N. (2013). The processes of inference. Argument and Computation, 4, 4–20.

Lane, D. S., Fletcher, D. N., & Fletcher, H. J. (1983). Improving conditional syllogism performance of young normal and gifted students with discovery and rule instruction. Journal of Educational Psychology, 75(3), 441–449. https://doi.org/10.1037/0022-0663.75.3.441

Le, N.-T., & Wartschinski, L. (2018). A Cognitive Assistant for improving human reasoning skills. International Journal of Human-Computer Studies, 117, 45–54. https://doi.org/10.1016/j.ijhcs.2018.02.005

Love, B. C. (2002). Comparing supervised and unsupervised category learning. Psychonomic Bulletin & Review, 9(4), 829–835.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to Information Retrieval (1 edition). Cambridge University Press.

McGee, V. (1985). A Counterexample to Modus Ponens. The Journal of Philosophy, 82(9), 462–471. https://doi.org/10.2307/2026276

Nückles, M., & Wittwer, J. (2014). Lernen und Wissenserwerb. In T. Seidel & A. Krapp (Hrsg.), Pädagogische Psychologie (S. 225-252). Weinheim: Beltz.

Page, E. B. (1963). Ordered Hypotheses for Multiple Treatments: A Significance Test for Linear Ranks. Journal of the American Statistical Association, 58(301), 216–230. https://doi.org/10.1080/01621459.1963.10500843

Piaget, J. (1936). Origins of intelligence in the child. London: Routledge & Kegan Paul.

Ragni, M., Eichhorn, C., Bock, T., Kern-Isberner, G., & Tse, A. P. P. (2017). Formal Nonmonotonic Theories and Properties of Human Defeasible Reasoning. Minds and Machines, 27(1), 79–117. https://doi.org/10.1007/s11023-016-9414-1

Ragni M., Riesterer, N., Khemlani, S., Johnson-Laird, P.N. (2018) How Stable is Human Reasoning? Manuscript submitted for publication.

Revlis, R. (1975). Two models of syllogistic reasoning: Feature selection and conversion. Journal of Verbal Learning and Verbal Behavior, 14(2), 180–195. https://doi.org/10.1016/S0022-5371(75)80064-8

Roessingh, J. J. M., Hilburn, B. G., Nationaal, N., & Ruimtevaartlaboratorium, L. (2000). NLR-TP-2000-308 The Power Law of Practice in adaptive training.

Russell, P. N. S. J. (2015). Artificial Intelligence: A Modern Approach, 3rd Edition. Pearson India.

Scouller, K. (1998). The influence of assessment method on students' learning approaches: Multiple choice question examination versus assignment essay. Higher Education, 35(4), 453–472. https://doi.org/10.1023/A:1003196224280

Simpson, J. A., Weiner, E. S. C., & Oxford University Press. (1989). The Oxford English Dictionary. Oxford; Oxford; New York: Clarendon Press; Oxford University Press.

Stenning, K., & Lambalgen, M. van. (2012). Human Reasoning and Cognitive Science (1. edition). Cambridge, Mass: A Bradford Book.

Sweller, J. (2008). Instructional Implications of David C. Geary's Evolutionary Educational Psychology. Educational Psychologist, 43(4), 214–216. https://doi.org/10.1080/00461520802392208

Taatgen, N. A., & Lee, F. J. (2003). Production Compilation: A Simple Mechanism to Model Complex Skill Acquisition, Production Compilation: A Simple Mechanism to Model Complex Skill Acquisition. Human Factors, 45(1), 61–76. https://doi.org/10.1518/hfes.45.1.61.27224

Tomasello, M., & Herrmann, E. (2010). Ape and Human Cognition: What's the Difference? Current Directions in Psychological Science, 19(1), 3–8. https://doi.org/10.1177/0963721409359300

Toplak, M. E., West, R. F., & Stanovich, K. E. (2011). The Cognitive Reflection Test as a predictor of performance on heuristics-and-biases tasks. Memory & Cognition, 39(7), 1275. https://doi.org/10.3758/s13421-011-0104-1

Wilkins C., (1928). The Effect of Changed Material on Ability to Do Formal Syllogistic Reasoning. Archives of Psychology, 16, 83.

Wilson, D. (2016). Relevance Theory. https://doi.org/10.1093/oxfordhb/9780199697960.013.25

Woodworth, R. S., & Sells, S. B. (1935). An atmosphere effect in formal syllogistic reasoning. Journal of Experimental Psychology, 18(4), 451–460. https://doi.org/10.1037/h0060520

Zadeh, L. A. (1997). Toward a theory of fuzzy information granulation and its centrality in human reasoning and fuzzy logic. Fuzzy Sets and Systems, 90(2), 111–127. https://doi.org/10.1016/S0165-0114(97)00077-8

APPENDIX A

Syllogism Tasks

1.) Premise 1: All linguists are carpenters. Premise 2: All carpenters are divers.

2.) Premise 1: All students are buyers. Premise 2: All cyclists are students.

3.) Premise 1: All linguists are divers. Premise 2: All carpenters are divers.

4.) Premise 1: All tutors are butchers. Premise 2: All tutors are packers.

5.) Premise 1: All pilots are singers. Premise 2: Some singers are tailors.

6.) Premise 1: All drillers are nurses. Premise 2: Some actors are drillers.

7.) Premise 1: All miners are poets. Premise 2: Some skaters are poets.

8.) Premise 1: All boxers are opticians. Premise 2: Some boxers are actuaries.

9.) Premise 1: All hikers are drivers. Premise 2: No drivers are writers.

10.) Premise 1: All carpenters are linguists. Premise 2: No divers are carpenters.

11.) Premise 1: All poets are waiters. Premise 2: No cashiers are waiters.

12.) Premise 1: All butchers are tutors. Premise 2: No butchers are packers.

13.) Premise 1: All divers are linguists. Premise 2: Some linguists are not carpenters.

14.) Premise 1: All auditors are cleaners. Premise 2: Some painters are not auditors.

15.) Premise 1: All skaters are miners. Premise 2: Some poets are not miners.

16.) Premise 1: All waiters are poets. Premise 2: Some waiters are not cashiers.

17.) Premise 1: Some sailors are potters. Premise 2: All potters are chemists.

18.) Premise 1: Some boxers are dancers. Premise 2: All typists are boxers.

19.) Premise 1: Some novelists are travelers. Premise 2: All analysts are travelers.

20.) Premise 1: Some novelists are travelers. Premise 2: All novelists are analysts.

21.) Premise 1: Some farmers are assistants. Premise 2: Some assistants are scholars.

22.) Premise 1: Some dentists are lifeguards. Premise 2: Some plumbers are dentists.

23.) Premise 1: Some painters are auditors. Premise 2: Some cleaners are auditors.

24.) Premise 1: Some surfers are artists. Premise 2: Some surfers are planners.

25.) Premise 1: Some tutors are packers. Premise 2: No packers are butchers.

26.) Premise 1: Some athletes are hunters. Premise 2: No lawyers are athletes.

27.) Premise 1: Some models are clerks. Premise 2: No managers are clerks.

28.) Premise 1: Some actors are drillers. Premise 2: No actors are nurses.

29.) Premise 1: Some plumbers are lifeguards. Premise 2: Some lifeguards are not dentists.

30.) Premise 1: Some cashiers are poets. Premise 2: Some waiters are not cashiers.

31.) Premise 1: Some gamblers are bakers. Premise 2: Some sculptors are not bakers.

32.) Premise 1: Some engineers are fencers. Premise 2: Some engineers are not campers.

33.) Premise 1: No fencers are engineers. Premise 2: All engineers are campers.

34.) Premise 1: No typists are dancers. Premise 2: All boxers are typists.

35.) Premise 1: No campers are fencers. Premise 2: All engineers are fencers.

36.) Premise 1: No engineers are fencers. Premise 2: All engineers are campers.

37.) Premise 1: No cooks are mayors. Premise 2: Some mayors are swimmers.

38.) Premise 1: No bakers are gamblers. Premise 2: Some sculptors are bakers.

39.) Premise 1: No agents are brokers. Premise 2: Some secretaries are brokers.

40.) Premise 1: No actors are painters. Premise 2: Some actors are workers.

41.) Premise 1: No planners are artists. Premise 2: No artists are surfers.

42.) Premise 1: No chefs are scientists. Premise 2: No runners are chefs.

43.) Premise 1: No judges are brewers. Premise 2: No porters are brewers.

44.) Premise 1: No actuaries are opticians. Premise 2: No actuaries are boxers.

45.) Premise 1: No climbers are doctors. Premise 2: Some doctors are not tellers.

46.) Premise 1: No soldiers are editors. Premise 2: Some florists are not soldiers.

47.) Premise 1: No surfers are planners. Premise 2: Some artists are not planners.

48.) Premise 1: No cleaners are auditors. Premise 2: Some cleaners are not painters.

49.) Premise 1: Some soldiers are not florists. Premise 2: All florists are editors.

50.) Premise 1: Some bankers are not teachers. Premise 2: All golfers are bankers.

51.) Premise 1: Some artists are not planners. Premise 2: All surfers are planners.

52.) Premise 1: Some jugglers are not investors. Premise 2: All jugglers are barbers.

53.) Premise 1: Some dentists are not lifeguards. Premise 2: Some lifeguards are plumbers.

54.) Premise 1: Some chemists are not potters. Premise 2: Some sailors are chemists.

55.) Premise 1: Some boxers are not typists. Premise 2: Some dancers are typists.

56.) Premise 1: Some soldiers are not editors. Premise 2: Some soldiers are florists.

57.) Premise 1: Some sculptors are not bakers. Premise 2: No bakers are gamblers.

58.) Premise 1: Some painters are not actors. Premise 2: No workers are painters.

59.) Premise 1: Some linguists are not carpenters. Premise 2: No divers are carpenters.

60.) Premise 1: Some plumbers are not dentists. Premise 2: No plumbers are lifeguards.

61.) Premise 1: Some counselors are not joggers. Premise 2: Some joggers are not riders.

62.) Premise 1: Some scientists are not chefs. Premise 2: Some runners are not scientists.

63.) Premise 1: Some barbers are not investors. Premise 2: Some jugglers are not investors.

64.) Premise 1: Some riders are not joggers. Premise 2: Some riders are not counselors.

APPENDIX B
Cognitive Reflection Tasks

a) A bat and a ball cost $1.10 in total. The bat costs a dollar more than the ball. How much does the ball cost?

b) If it takes 5 machines 5 minutes to make 5 widgets, how long would it take 100 machines to make 100 widgets?

c) In a lake, there is a patch of lily pads. Every day, the patch doubles in size. If it takes 48 days for the patch to cover the entire lake, how long would it take for the patch to cover half of the lake?

d) If John can drink one barrel of water in 6 days, and Mary can drink one barrel of water in 12 days, how long would it take them to drink one barrel of water together?

e) Jerry received both the 15th highest and the 15th lowest mark in the class. How many students are in the class?

f) A man buys a pig for $60, sells it for $70, buys it back for $80, and sells it finally for $90. How much has he made?

g) Simon decided to invest $8,000 in the stock market one day early in 2008. Six months after he invested, on July 17, the stocks he had purchased were down 50%. Fortunately for Simon, from July 17 to October 17, the stocks he had purchased went up 75%. At this point, Simon has: