

UTRECHT UNIVERSITY

THE ROLE OF CONTEXT AND SEMANTICS IN
REASONING

UNDERSTANDING THE NORMATIVE/DESCRIPTIVE GAP

Stamatis Kantiloros

Supervisor
Dr. Colin Caret

Second Examiner
Dr. Chris Janssen



Universiteit Utrecht

Graduate School of Natural Sciences
Artificial Intelligence
August 2020

Abstract

It has been found through experimentation, that often people can suppress logically valid inferences. When additional or alternative information is added to a combination of premises depicting the same kind of inference that was at first thought logically valid, people reason to different conclusions in a manner that seems irrational. The Suppression task is a practical example in which one can observe the divergence between human thinking and first-order Logic. In this research, a within-subject design experiment was devised in a form of a simple questionnaire that was distributed online through Google Forms. The experiment focused on exposing the role of context and semantics in reasoning, especially in syllogistic tasks that involve mathematical inferences. The experimental results are reported, along with some exploratory analysis and an extensive literature review on the main causes of the Normative/Descriptive gap.

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 4 |
| 1.1 | Artificial Intelligence and Human Cognition | 4 |
| 1.2 | Normative, Descriptive and Prescriptive Models | 5 |
| 1.3 | The Divergence from Normative Models | 7 |
| 1.4 | The Relevance of Logic | 7 |
| 2 | Overview of the most common explanations of irrationality | 9 |
| 2.1 | Algorithmic Level Limitations and Performance Errors | 9 |
| 2.1.1 | Performance Errors | 9 |
| 2.1.2 | Algorithmic Level Limitations | 11 |
| 2.2 | Alternative Problem Construal | 13 |
| 2.2.1 | The Process of Interpretation | 14 |
| 2.2.2 | Empirical Evidence | 15 |
| 2.3 | Incorrect Norm Application | 16 |
| 2.4 | Systematic Irrationality in the Intentional-Level Psychology | 17 |
| 2.5 | The Need for Normative Models | 18 |
| 3 | The Suppression Task | 20 |
| 3.1 | The Suppression Task Experiment | 20 |
| 3.2 | Original Experiment | 20 |
| 3.2.1 | Interpretation and experimental procedures | 23 |
| 3.2.2 | Confounds of the original experiment | 23 |
| 3.2.3 | Interpretation of the results | 24 |
| 4 | Experimental Design | 25 |
| 4.1 | Framing | 25 |
| 4.2 | Experimental Design | 26 |
| 4.2.1 | Research Question | 26 |
| 4.3 | Method | 27 |
| 4.3.1 | Participants | 27 |
| 4.3.2 | Materials Used | 27 |
| 4.3.3 | Procedure | 30 |
| 4.3.4 | Design | 30 |
| 4.3.5 | Measures | 30 |
| 4.4 | Expectations | 31 |
| 5 | Results | 32 |
| 5.1 | Statistics | 32 |
| 5.1.1 | Inferential Statistics | 33 |
| 5.2 | Discussion on Experimental Results | 34 |
| 5.2.1 | Feedback from Participants | 34 |

| | |
|---|-----------|
| 6 General Discussion | 36 |
| 6.1 Criticisms | 36 |
| 6.2 Exploratory Analysis | 38 |
| 6.3 Relation to The Most Common Explanations of Irrationality . . | 39 |
| 6.4 Closing Remarks | 40 |
| Appendices | 41 |
| References | 49 |

1 Introduction

Most people would probably agree with the proposition, that, modeling human cognition and Artificial Intelligence (AI) are highly related to each other. This study aims to make a small contribution towards informing models of human cognition and aid in making them more accurate and realistic.

1.1 Artificial Intelligence and Human Cognition

Generally, when thinking about what AI really is, the words of philosopher John Haugeland come to mind.

“The exciting new effort to make computers think ... machines with minds, in the full and literal sense.” (Haugeland, 1985)

Inspiring as these words may sound, accurately defining AI has proven to be an insurmountable task. And that is because a consensus definition of AI depends on first assigning definitions to a large set of highly contentious and ambiguous concepts. The reader, especially one that might have taken classes in the area of Philosophy, can appreciate the severity of the problem just by thinking about the following questions: What is intelligence? What is consciousness? Who is rational? Is there free will?

In order to bypass having to answer all these difficult questions, we can at least give some proposed working definitions that focus on defining AI in terms of its goals. Such definitions were given by Russel and Norvig in their book “Artificial Intelligence - A modern Approach” (Table 1). As we can observe, a lot of it has to do with setting points of reference. On the one hand, we have a human-centric approach that has the goal of accurately mimicking human behavior and performance. On the other hand, we have an intelligence maximalism approach that aims to achieve absolute optimum in a system in terms of reasoning and goal-oriented behavior.

| | Human-Based | Ideal Rationality |
|------------------------|---------------------------------|--------------------------------|
| Reasoning Based | Systems that think like humans. | Systems that think rationally. |
| Behavior Based | Systems that act like humans. | Systems that act rationally. |

Table 1: Four Possible Goals for AI (Russell & Norvig, 2002)

As far as what cognition actually is, we can look at the Oxford Dictionary of Psychology to get a quick answer.

“The mental activities involved in acquiring and processing information” (Stevenson, 2010)

So, it is safe to say that human cognition entails acquiring and processing information by a human.

Sometimes, people think AI is a more engineering oriented field, whereas cognitive science is purely theoretical. In some cases that is true, but when we

consider the big challenges each field faces, then usually there is a convergence between them.

Let us, for example, consider the problem of building an autonomous vehicle. Indeed, one could argue that a general solution for controlling the behavior of the vehicle does not necessarily have any consequences for human cognition. However, that speaks more to the cognitive potency of humans rather than the cognitive inability of the vehicle. In other words, the vehicle’s cognitive ability, or in general the cognitive ability of any robotic system, is a small subset of human cognition. But, this is not because the designers of such systems made a conscious choice to “dumb” them down. They would very much like to have built systems that are as intelligent as humans, or, even more. In such a case, then, a general cognitive model of human cognition would be an ideal solution for this AI engineering problem, which at the same time would be a milestone achievement for cognitive science.

In the grand context of science, AI and human cognition operate under a lot of similar constraints. They both have as a mandate to acquire and process information in order to achieve some objective. They only differ in the manner in which the acquisition and processing of information takes place, as well as the overall goals that they are trying to achieve by performing these activities. Now, these are no small differences by any means, but there is a lot of common ground between them in order to render advances in either field mutually beneficial to each other.

1.2 Normative, Descriptive and Prescriptive Models

Trying to accurately model human cognition has proven to be one of the most difficult tasks researchers across many fields of science have engaged with. Generally, we can recognize three major categories of cognitive models that have been developed: *Normative*, *Descriptive* and *Prescriptive* models (Stanovich, 1999).

Normative models define analytical frameworks that contain rules and constraints on what constitutes good reasoning. They tell us whether an agent is reasoning correctly, according to some set of goals.

Normative models, as noted, are standards for evaluation. They must be justified independently of observations of people’s judgments and decisions, once we have observed enough to define what we are talking about. (Baron, 2012)

Descriptive models aim to capture a theoretical account of the response patterns of human thinkers.

Descriptive models are psychological theories that try to explain how people make judgments and decisions, typically in the language of cognitive psychology, which includes such concepts as heuristics and strategies, as well as formal mathematical models. (Baron, 2012)

Prescriptive models aim to create methods that improve reasoning performance in situations when a theoretically optimal response is not possible due to certain limitations.

Prescriptive models are designs for improvement. If normative models fall in the domain of philosophy (broadly defined) and descriptive models in the domain of empirical psychological science, then prescriptive models are in the domain of engineering (again, broadly defined). (Baron, 2012)

On the one hand, some researchers can argue that normative criteria should be the epicenter of a framework that aims to model human cognition and psychology can only have a secondary, descriptive role. On the other hand, there are many well founded arguments that dispute the absolute role of normative rules in modeling human cognition and argue that empirical findings should play a vital role in informing the overall model. Obviously since no universal model of human cognition that is accurate and widely accepted exists, the makeup of such a model is open for discussion.

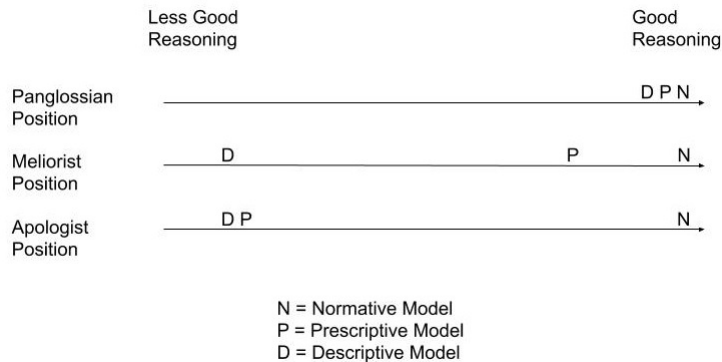


Figure 1: Three Theoretical Positions on Human Rationality

Moreover, the debate between scientists over which kind of model accurately captures human reasoning has produced 3 large schools of thought, namely the Meliorists, the Apologists and the Panglossians.

The *Panglossian* position sees no gaps between the descriptive and the normative. Since there are no gaps, human behavior is deemed to be, largely, rational. In instances where human behavior does differ from the normative one, the departure is attributed to performance errors, minor cognitive slips, memory lapses or other unimportant psychological malfunctions. Additionally, it can be argued that the experimenter is applying the wrong normative model to a certain cognitive problem or that the subject has a different construal of the cognitive task at hand.

The *Meliorist* position states that a prescriptive model is quite close to a normative one. However, it accepts that human behavior can depart significantly from the normative. Because the prescriptive is thought to be close to

the normative, actual behavior can be quite far from the optimal computable response. Therefore, the Meliorist position leaves substantial room for improvement in human reasoning when it is possible to bridge the gap between the descriptive and the prescriptive.

The *Apologist* position accepts that there might be a large gap between the descriptive and the normative, and sees that a prescriptive model is an accurate depiction of how people usually reason. According to this view, irrationality might not be easily attributed to human reasoning as computational limitations and lack of knowledge from the part of human thinkers should be taken seriously.

These schools of thought provide us with a point of reference along with convenient structure and labeling of various arguments regarding human cognition.

1.3 The Divergence from Normative Models

Trying to imagine how a perfect model of human cognition would be constructed though is, on its own, a gigantic and almost impossible task. But we can narrow down our focus to one of the most interesting occasions that one encounters when investigating the bodies of work in human cognition: when human thinking deviates from the kind of thinking that would be deemed correct according to various normative models.

Examples of such behavior can include people violating axioms of utility theory, misjudge the probability of potential outcomes or fail to adhere to simple principles of Logic (Stanovich & West, 1998). In the literature of reasoning and decision making, the phenomenon in which human performance deviates from the behavior considered normative under certain models of optimal response, is called the Normative/Descriptive gap.

Broadly speaking we can identify 5 explanations that are mentioned throughout relevant literature for the Normative/Descriptive gap (Stanovich, 1999).

- Systematic irrationality in the intentional-level psychology
- Algorithmic level limitations
- Performance errors
- Incorrect norm application
- Alternative problem construal

1.4 The Relevance of Logic

In a way Logic offers the perfect normative model, when it is appropriate to assume that Logic represents a normative model for a certain cognitive task. Therefore, in such cases, it should be easier to at least dispute the “incorrect norm application” argument.

However, due to the constrained nature of the way human cognition happens in real life, pure Logic may not always be a reasonable competence model. Instead, we can consider a form of non-monotonic logic as a competence model,

namely closed-world reasoning. We should also note here that Closed-world reasoning can also be affected by situational parameters like context and semantics (Stenning & Van Lambalgen, 2012).

Whether we assume Logic or indeed closed-world reasoning as the perfect normative models for certain cognitive tasks, the work of a researcher is not done as the other 4 possible explanations of the normative/descriptive gap that were motioned previously still need to be addressed. And there needs to be a thorough explanation as to why Logic or closed-world reasoning were chosen as the normative models for a certain cognitive task.

All in all, demonstrating discrepancies between descriptive accounts of behavior and normative models is not anything novel in this line of research. But, the theoretical interpretation of these discrepancies is still highly contentious and provides a fertile ground for further exploration.

The Suppression Task experiment (Byrne, 1989) is such an occasion, where the normative model is assumed to be Logic, and we observe human subjects diverging from this model. It is therefore convenient to use it as a case study for this work.

2 Overview of the most common explanations of irrationality

Since this study is focusing on instances of irrationality, it is beneficial to first go over the most common explanations of irrationality that can be found in relevant literature.

2.1 Algorithmic Level Limitations and Performance Errors

2.1.1 Performance Errors

Sometimes discrepancies between actual responses and those dictated by normative models are attributed to performance errors. For a precise definition of what is meant by “performance errors”, we cite Stein, who was amongst the first to accurately describe what this concept involves (Stein, 1996).

A momentary lapse, a divergence from some typical behavior. This is in contrast to attributing a divergence from norm to reasoning in accordance with principles that diverge from the normative principles of reasoning. Behavior due to irrationality connotes a systematic divergence from the norm. It is this distinction between mere mistakes and systematic violations (between performance errors and competence errors) that is ... implicitly assumed by friends of the rationality thesis when they deny that the reasoning experiments [demonstrate human irrationality]. (p. 8)

It is important to mention here that an adoption of a “strong” version of this view would have it so that virtually all observed departures from normative models are explained by performance errors. Is it possible to find evidence that justify this view?

There have been numerous studies that show that such a strong view cannot be supported. The main argument researchers use here is that, if each instance of irrationality represents a momentary processing lapse due to distraction or confusion, well then we wouldn’t expect to observe any kind of correlation in human performances across different cognitive tasks. Rips and Conrad’s studies demonstrate this point through experimental data.

“Subjects absolute scores on the propositional tests correlated with their performance on certain other reasoning tests. ...If the differences in propositional reasoning were merely due to interference from other performance factors, it would be difficult to explain why they correlate with these tests” (p. 282-283) (Rips & Conrad, 1983).

And in more recent years, extensive studies done by Stanovich and West (Stanovich & West, 1998) that examined a variety of tasks from the heuristics and biases literature, showed significant cross-task correlations where subjects that gave a normative response on a certain task were also more likely to give it on another.

A very “weak” version though of the performance error argument that challenges its potency to explain divergences from normative behavior is also not likely (Bell, Gardner, & Woltz, 1997).

In more recent years, there has been efforts to incorporate the modeling of human errors in a more concise manner. These researchers defined a formal behavioral model of human errors (BMHE) based on human and system behaviors that aims to explain the fundamental mechanisms of human errors.

Definition: *A human behavior B is constituted by four basic elements known as the sets of objects (O), actions (A), space (S), and time (T):*

$$B = (O,A,S,T) = O \times A \times S \times T \text{ (Wang, 2008)}$$

Any incorrect configuration of any of these four elements results in a human error in task performance. This notion, combined with the systematic human error reduction and prediction approach (SHERPA) that was proposed by D.Embry in 1986 outputs the table of BMHE(Figure 2).

| No. | Objects | Action | Space | Time | Error Mode |
|-----|---------|--------|-------|------|---|
| 0 | T | T | T | T | Correct action |
| 1 | T | T | T | F | Mode 1: Wrong timing |
| 2 | T | T | F | T | Mode 2: Wrong place |
| 3 | T | T | F | F | Mode 3: Wrong timing and place |
| 4 | T | F | T | T | Mode 4: Wrong action |
| 5 | T | F | T | F | Mode 5: Wrong action and timing |
| 6 | T | F | F | T | Mode 6: Wrong action and place |
| 7 | T | F | F | F | Mode 7: Wrong action, place, and timing |
| 8 | F | T | T | T | Mode 8: Wrong object |
| 9 | F | T | T | F | Mode 9: Wrong object and timing |
| 10 | F | T | F | T | Mode 10: Wrong object and place |
| 11 | F | T | F | F | Mode 11: Wrong object, place, and timing |
| 12 | F | F | T | T | Mode 12: Wrong object and action |
| 13 | F | F | T | F | Mode 13: Wrong object, action, and timing |
| 14 | F | F | F | T | Mode 14: Wrong object, action, and place |
| 15 | F | F | F | F | Mode 15: All wrong |

Figure 2: The Behavioral Model of Human Errors(Wang, 2008)

While such attempts at trying to accurately model human errors are aiming at the right direction, they usually fall sort, because of the highly complex variance of tasks humans perform. This means that they usually operate under severe constraints and rapidly fall apart when they are tested in a wide variety of tasks.

It can be argued, of course, in defense of those models, that they are optimized for a specific domain, in which they perform quite well. That can certainly

be the case, but then we have to clarify that these models are domain specific, and they are also dependent on the choice of an appropriate normative model for the task at hand. So in the end, the non-generalizability of these models takes away a lot of their potency as tools that accurately describe human behavior in a wide variety of situations.

What is probably more useful, though, in this study is the description of the statistical properties of human errors that these researchers have to offer. They attribute oddness, independence and randomness as core features of human performance errors (Wang, 2008).

Oddness, means that, counter intuitively, although individuals make different errors while performing tasks, there is a higher chance of making a single error in a specific task than that of making multiple errors.

Independence, means different individuals can have different error patterns while performing the same task.

Randomness, means different individuals can make the same error but at a different place when performing a task.

So, where do performance errors sit in the grand scheme of things as a possible explanation of human irrationality? Their odd and random nature makes it difficult to model them, so that we can predict them with a relatively high amount of accuracy. It seems like performance errors are a native characteristic of humans and they exist as a reminder of the imperfect nature of human reasoning. However they hold little to no explanatory value in terms of defining and modeling a cause for systematic divergences between normative and descriptive models. As such, they will always introduce randomness and noise when one observes human behavioral data.

2.1.2 Algorithmic Level Limitations

Algorithmic level limitations are one of the main reasons that justify the need for prescriptive models, along with environmental limitations (limited time, space etc.). Accurately accessing them is quite a challenge that usually requires empirical work, but in return provides us with valuable insights on the constrained nature of human cognition.

Here we examine the case where human performance would be very close or at normative levels, however computational limitations are the main hindrance that create the normative/descriptive gap. To classify behavior as irrational via comparison with a normative model, we would have to account for the computability of said normative model in the context of the limited cognitive capacity of human beings.

But, accounting for the limited nature of cognitive human capabilities is no easy task. We would have to come up with a way of defining and measuring human cognitive capacity. Then, even if that endeavour is successful we would need to tackle the issue of individual differences between humans.

It has been found that there is a strong association between cognitive ability and task performance, and that this association can be used as an indicator of human cognitive capacity (Stanovich, 1999). Stanovich and West used a wide

variety of tasks such as the Stochastic Aptitude Test(SAT), Raven Matrices and various vocabulary and reading comprehension tests.

| Correlations Between Performance on the Reasoning Tasks and SAT Total Score | |
|--|--------|
| <i>Data from Study 1 of Stanovich and West (1998b)</i> | |
| Syllogisms | .470* |
| Selection task | .394* |
| Statistical reasoning | .347* |
| Argument evaluation | .358* |
| <i>Replication and Extension (Study 2 of Stanovich & West, 1998b)</i> | |
| Syllogisms | .410* |
| Statistical reasoning | .376* |
| Argument evaluation task | .371* |
| Covariation detection | .239* |
| Hypothesis-testing bias | -.223* |
| Outcome bias | -.172* |
| If/only thinking | -.208* |
| RT1 composite | .530* |
| RT2 composite | .383* |
| RT composite, all tasks | .547* |

Note. RT1 composite = standard score composite of performance on argument evaluation task, syllogisms, and statistical reasoning; RT2 composite = standard score composite of performance on covariation judgment, hypothesis-testing task, if/only thinking, and outcome bias; RT composite, all tasks = rational thinking composite score of performance on all seven tasks in the replication and extension experiment.

* = $p < .001$, all two-tailed. N s = 527 to 529 in the replication and extension.

Figure 3: (Stanovich & West, 1998)

A reasonable objection to this finding is that this approach does not account for different educational backgrounds amongst individuals. A stronger and more diverse education could obviously provide an explanation as to why the individual who possess it can give normatively appropriate responses to cognitive tasks. When the researchers accounted for this fact they found that across seven tasks, people with a mathematics/statistics background performed significantly better only in one of them (Argument Evaluation).

As it can be seen in Figure 4, the composite variable of all seven thinking tasks displayed a correlation of only 0.162 which is much lower than that with the SAT Total Score which is 0.547 (Figure 3).

Therefore it seems that systematic discrepancies between actual performance and normative models can be accounted for by variation in cognitive abilities, to a moderate extent. In the studies mentioned above, individuals

**Correlations Between Performance on the Reasoning Tasks and
Mathematics/Statistics Background**

| <i>Mathematics/Statistics Background</i> | |
|--|--------|
| Argument evaluation task | .137* |
| Syllogisms | .091 |
| Statistical reasoning | .075 |
| Covariation detection | .088 |
| Hypothesis-testing bias | -.045 |
| Outcome bias | -.071 |
| If/only thinking | -.062 |
| RT1 composite | .145* |
| RT2 composite | .125* |
| RT composite, all tasks | .162** |

Figure 4: (Stanovich & West, 1998)

that give a normative response to one task are more likely to also give a normative response to a different task as well.

So, we conclude that algorithmic level limitations can indeed offer a potent explanation, that is widely agreed upon in relevant literature, for differences that arise between normative and descriptive models. It also offers fertile ground for researching the individualities amongst humans when it comes to reasoning.

2.2 Alternative Problem Construal

In this case we accept that the researcher has indeed chosen the correct normative model, but the subjects are construing the task differently and therefore providing a normatively appropriate answer to a different problem. So there must be features of the experimental task that lead subjects to interpret the problem in a different manner. This issue, then, places the blame partly on the experimenter for deploying a poor experimental design that can potentially mislead the subjects.

As always, evaluating human behavior is a delicate subject. Because, were we to adopt a “strong” view of the Alternative Problem Construal argument, we would tend to reject normative models and their explanatory capabilities for human cognitive behavior. But, the position that is going to be adopted here is that normative appropriateness of a response to a particular task is always relative to a particular interpretation of the task.

If interpretation plays a crucial role, then we have to find a way to systematically think about it. Otherwise it will always present itself as a blank space and potential hidden experimental confound.

2.2.1 The Process of Interpretation

The “translation” from natural language to a formal logic-like language is not a trivial procedure. This can challenge the efficacy and appropriateness of a normative model, as it is natural to expect that if we apply the same norms to different problems, we can expect to get different solutions. Stenning and Van Lambalgen propose a framework in which reasoning happens in two important and distinct stages (Stenning & Van Lambalgen, 2012). The framework underlines the importance of this “translation”, which really is the process of interpretation.

- First, the domain in which one reasons needs to be established along with its formal properties. This is what will be known as reasoning *to* an interpretation.
- Second, the formal laws that reasoning needs to adhere to are specified. This is what will be known as reasoning *from* an interpretation.

Using the previously established guidelines, the process of interpretation will at least need to account for two variables: what things exist in the domain we need to reason in and what kind of reasoning will be done about them. Let’s examine this paragraph which is also featured in Stenning and Van Lambalgen’s book.

Once upon a time there was a butcher, a baker, and a candlestick maker. One fine morning, a body was discovered on the village green, a dagger protruding from its chest. The murderer’s footprints were clearly registered in the mud...
(Stenning & Van Lambalgen, 2012)

How does one determine in full detail the domain in which the story unfolds is not entirely clear. We are told that the events that are described take place in a village. However we do not know if this is the only village in the entire story of the book, or there are many villages, or, in fact, this is a village that belongs to a country and there are many countries with many villages, towns etc. So let us imagine ourselves in the role of a crime investigator for this scene. We need to answer the question: “Where do the events take place?”

If we assume that the village is the only geographical location that exists in the world of the book, that is to say the entire world consists of just this village, then we know the murderer can only come from one place.

If we assume that the village is part of a country with many towns and villages, well then we have to consider that the murderer could come from this village or some other town or village.

If we assume that there are many countries with many villages and towns, yet again we have to widen the scope of the possibilities for our search to be accurate.

What this demonstrates is that the process of interpretation is not limited to assigning meaning to lexical terms. There is a deeper layer in the contextual information that plays a crucial role in determining the domain in which the story operates. Any time interpretation of incoming information is considered then, the recipient of such information has to develop a judgement on its source. This will help in defining more accurately the context in which the information is supposed to be situated.

This judgement, broadly speaking, can either err on the side of trustworthiness, or on the side of skepticism. In relevant literature, usually those two opposing views are mapped to two different kinds of reasoning: *credulous* and *skeptical* reasoning.

Credulous reasoning makes the assumption that the speaker's utterances or the author's excerpts are true. Therefore the source of information is trustworthy. So when we encounter contradictions or dead-ends of incomplete information we should aim to modify our interpretation in order to restore the validity of the model of discourse we have constructed.

Skeptical reasoning makes the assumption that the source of information should never be blindly trusted. Therefore if contradictions arise we can challenge the information that is being presented to us on the basis that a conclusion might be false under a certain interpretation of the premises.

It is fair to claim that people deploy *credulous* and *skeptical* reasoning in varying degrees in everyday life. Therefore we would expect the same to happen in an experimental setting as well, a fact that can directly influence the perception of a cognitive task.

2.2.2 Empirical Evidence

Whether it is *credulous* reasoning, *skeptical* reasoning or some other aspect of the interpretation process, it has been articulated that the perception of a cognitive task is not a straightforward process. Even worse, evaluating alternative problem construals is a daunting task as it is incredibly difficult to assess what other task might have been constructed in a subject's mind.

Looking at relevant literature, we can observe studies that have been done using popular tasks that act as a point of reference such as the Linda Problem (Tversky & Kahneman, 1983), the Disease Problem (Tversky & Kahneman, 1981), and the selection task (Wason, 1966). Going over those studies, in his book Stanovich showed that more often than not the original problem construal, as it was designed by the experimenters, was favored by people of high analytic intelligence. In contrast, alternative construals were favored by individuals of lower analytic intelligence. He also argues about the possibility that alternative construals may be triggered by heuristics that make evolutionary sense, although subjects that possess higher and more flexible analytic intelligence are more likely to follow normative rules that maximize personal utility (Stanovich, 1999).

So, the Alternative Problem Construal thesis can hold a lot of explanatory power as far as the normative/descriptive gap is concerned. But, there is a lot of fuzziness that comes with this approach, which in the end weakens its potency through imprecision. It is fairly easy to claim that a subject is misinterpreting a task during an experimental procedure. It is extremely difficult to accurately define the alternative interpretation that the subject has come to.

2.3 Incorrect Norm Application

This is a contentious topic by nature as it is easy to argue for or against a specific normative model for a certain task. It also underlines an important topic that is greatly highlighted in this work, which is that an experiment says as much about the subjects undertaking it as about the experimenters that designed it.

The origins of this problem can be traced to the fact that the field of Psychology more often than not, uses normative models from other disciplines to evaluate human behavior. Leaning heavily towards rejecting normativity would have us align with Panglossians, as we would have to accept that in this case people are indeed principally rational and our normative models are not properly constructed. On the other hand, claiming that people display systematic irrationality when they fail to give normative responses would have us align more with Meliorists while simultaneously supporting the soundness of our normative model.

There is a plethora of examples in literature where scientists critique the use of a certain normative model.

Wetherick while criticising such experimental paradigms writes

“What Wason and his successors judged to be the wrong response is in fact correct.” (Wetherick, 1995)

Messer and Griggs note in their book discuss another view on the classic Linda conjunction problem where they reject the notion that the problem itself constitutes a regular probability problem.

“The results are discussed in terms of Gigerenzer’s (1991) normative-issues argument that the Linda problem is not a frequency probability problem but rather a single-trial, subjective probability problem.” (Messer & Griggs, 1993)

Inspired partly by the same problem, Ralph Hertwig and Gerd Gigerenzer also make notes on the misinterpretation of the experimental tasks from the researchers that construct them.

“We conclude that a failure to recognize the human capacity for semantic and pragmatic inference can lead rational responses to be misclassified as fallacies.” (Hertwig & Gigerenzer, 1999)

So how should we think in situations where there are large discrepancies between normative and descriptive responses, and there are doubts about the chosen normative model?

Sadly, there is no easy answer to this question. Each instance of this case will vary and we will always have to dig deeper in order to understand in each

case what other factors might be causing the variance in responses.

2.4 Systematic Irrationality in the Intentional-Level Psychology

It is hard to imagine that those who think that human thinking is rational in principle, will accept the thesis that there can be evidence for intentional systematic irrationality. Nevertheless, modern psychology has increasingly paid attention to concepts that reside in a grey area between cognition and personality. Those concepts are usually referred to as *Thinking Dispositions*.

“Despite this diversity of terminology, most authors use such terms similarly to refer to relatively stable psychological mechanisms and strategies that tend to generate characteristic behavioral tendencies and tactics.” (Stanovich, 1999)

Thinking Dispositions differ from pure cognitive capacity and therefore can potentially provide alternative explanations for the normative/descriptive gap. They usually act as an index of individual differences in human reasoning and inform us about an individual’s goals and epistemic values. As such, they can signal the existence of systematically sub-optimal systems in the intentional level psychology.

| | | CCTDI-UK Analy. | CCTDI-UK Inquis. | CCTDI-UK Open. | CCTDI-UK Sys. | CCTDI-UK SeCon. | CCTDI-UK TrSe. |
|-------------------------|-----------------|--------------------|---------------------|-------------------|------------------|--------------------|-------------------|
| CCTST-UK Total Score | r | .130 | .097 | .011 | .012 | .180(*) | .293(**) |
| | Sig. (2-tailed) | .113 | .237 | .894 | .883 | .028 | .000 |
| | N | 150 | 150 | 150 | 150 | 150 | 150 |
| CCTST-UK Evaluation | r | .130 | .040 | .002 | .024 | .189(*) | .245(**) |
| | Sig. (2-tailed) | .114 | .628 | .984 | .770 | .021 | .003 |
| | N | 150 | 150 | 150 | 150 | 150 | 150 |
| CCTST-UK Inference | r | .087 | .131 | .022 | .006 | .109 | .250(**) |
| | Sig. (2-tailed) | .289 | .109 | .789 | .943 | .183 | .002 |
| | N | 150 | 150 | 150 | 150 | 150 | 150 |

** Correlation is significant at the 0.01 level (2-tailed).

* Correlation is significant at the 0.05 level (2-tailed).

Figure 5: Correlations between critical thinking dispositions and critical thinking skills (O’Hare, 2004).

This study investigated critical thinking amongst undergraduate students (O’Hare, 2004). The main conclusions of the thesis were that these tests had

significant potential for predicting degree attainment through thinking dispositions.

Then, we can look at Stanovich and West, whose work has been widely referenced here, for individualities in human reasoning (Stanovich & West, 1998). They conducted 4 studies, involving a variety of tasks from the heuristics and biases literature including the selection task, belief bias in syllogistic reasoning, argument evaluation, base-rate use, covariation detection, hypothesis testing, outcome bias, if-only thinking, knowledge calibration, hindsight bias, and the false consensus paradigm. They explored the extent to which cognitive ability and thinking dispositions can predict discrepancies between normative and descriptive models. They concluded that inclination toward certain suppositions can predict individual differences in performance on the cognitive tasks that were mentioned above.

2.5 The Need for Normative Models

Normativity is a uniquely human design, as humans are the only species that can create novel norms from scratch (Elqayam & Over, 2016). Although Panglossianism, Meliorism and Apologism differ in the views they hold for human rationality, one thing amongst them is common: the need for normative models to act as a point of reference.

In literature that is relevant to the exploration of human cognition one often encounters the *rationality paradox*. The rationality paradox stems from the fact that although humans appear to be highly intelligent they also simultaneously show evidence of numerous errors and biases in their thinking and reasoning when measured against normative standards associated with formal, logical systems or probability theory (Stupple & Ball, 2014). Even though this paradox seems straightforward, within it we can see a lot of ambiguity. What is an error or a bias in reasoning? Why should we compare human thinking to normative standards? And even if we do, what should those standards be? Obviously under the rationality paradox paradigm, we assume that there are descriptions of what human thinking *'is'* against prescriptions of what human thinking *'ought'* to be.

However the *'is-ought'* assumption of the rationality paradox has not gone uncontested, and rightly so. Prominent researchers like Elqyam and Evans have made the case that normativism "...should be strictly avoided given that the dubious is-ought inference that it invokes fosters misunderstandings and obstructs sound theorizing" (Elqayam & Evans, 2011). Most of the push-back against normativity in recent years can be summarized in the debate between normativism versus relativism. A Panglossian normativist would have to argue that all errors in reasoning are non systematic and can be explained by some version of a performance error or alternative problem construal. On the other hand, a strong relativist would argue that there are no universal normative standards to judge inferences and therefore there is no case for human thinking being principally irrational.

Both normativists and relativists are in essence arguing that we should think

of human reasoning as principally rational. They only differ in the way their justify this common position. But their difference is substantial. One equates actual human behavior with normative models and the other denies the very existence of normativity.

So, it is evident that both approaches are somewhat of a dead end. Partly because of their extremity and mainly because their arguments have been greatly challenged both empirically and theoretically. A moderate consideration of both arguments can lead to the formulation of a more practical thesis that finds the use of normative standards beneficial for research and analytical purposes as long as they are used sensibly and with an understanding of their limitations.

More interestingly, we can look introspectively at the line of work that is displayed here, or for that matter, any thesis, paper or research that adhere to the *Scientific Method*. Although many fields have their own specific guidelines under which Scientific Method operates, it would be virtually impossible to carry out research on a large scale without the use of general guidelines. Generally scientists formulate hypotheses and test them by carrying out experiments or laying out theoretical argumentation that either supports or falsifies them.

The need for normative models, in the end, comes down to precision and the scientific method. All attempts to abolish them ultimately fail, because all other alternatives are too vague and general to be useful and withstand rigorous criticisms. Normative models have the ability to provide frameworks for reasoning, decision making and argumentation with a level of precision and completeness that can not be matched by descriptive or process-level analysis. Their efficacy as perfect models of human cognition can be very much disputed. But their potency as research tools and evaluation standards is unparalleled. So, their role is quite central to applied psychology and it would be hard to imagine what the field of psychology would be without them. In the following excerpt Baron illustrates this point eloquently.

“JDM is applied psychology. The ultimate goal is to improve judgments and decisions, or keep them from getting worse. In order to achieve this goal we need to know what good judgments and decisions are. That is, we need criteria for evaluation, so that we can gather data on the goodness of judgments, find out what makes them better or worse, and test method for improving them when there is room for improvement. This is the main function of normative models.”
(Baron, 2012)

3 The Suppression Task

3.1 The Suppression Task Experiment

It has been found through experimentation, that often people can suppress logically valid inferences. When additional or alternative information is added to a combination of premises depicting the same kind of inference that was at first thought logically valid, people reason to different conclusions in a manner that seems irrational.

The Suppression task is a practical example in which one can observe the divergence between human thinking and normative models. In this case, more specifically, human thinking constitutes of performing a logical task and the normative model is assumed to be first-order Logic.

3.2 Original Experiment

In the original experiment (Byrne, 1989), subjects were divided into three groups. One group received the conditional arguments with the *standard* antecedent, the second group received conditional arguments with an *alternative* antecedent and the third received conditional arguments with an *additional* antecedent.

In total there were 4 different subtasks that account for 4 different argument forms, Modus Ponens (MP), Modus Tollens (MT), Denial of the antecedent (DA), Affirmation of the consequent (AC). These subtasks, together with a combination of different premises account for the full Suppression Task experiment. Below can be found a table showing the results of the experiment of Dieussaert's subjects (Dieussaert, Schaeken, Schroyens, & d'Ydewalle, 2000).

Modus Ponens (MP). The premises (simple or otherwise), are given in the combination with proposition p: she has an essay to write. Participants are asked whether q (she will study late in the library) is true.

Modus Tollens (MT). The premises and the proposition $\neg q$ are given: she will not study late in the library. Participants are asked whether $\neg p$ (she does not have an essay to write) is true.

Denial of the antecedent (DA). The premises and the proposition $\neg p$ are given: she does not have an essay to write. Participants are asked whether q (she will not study late in the library) is true.

Affirmation of the consequent (AC). The premises and proposition q are given: she will study late in the library. Participants are asked whether p (she has an essay to write) is true.

All subjects received all four variations of the conditional arguments, modus ponens, modus tollens, denial of the antecedent and affirmation of the consequent. Each sort of argument was presented with three different contents and therefore each subject was presented with 12 total arguments in random order.

For reference, here is some of the exact instructions that were given to the subjects (Byrne, 1989).

For all subjects these instructions explained the task with reference to a simple argument as an example. They were asked to assume that the premises were true and to "choose one of the conclusions-(a),(b) or (c)- whichever you think follows from the sentences." Subjects were asked to read each item carefully and to work from beginning to end at their own pace, without changing any responses or skipping any items.

The subjects were presented with the sentences like so

If she has an essay to write, she will study late in the library.

She has an essay to write.

- (a) She will study late in the library.
- (b) She will not study late in the library.
- (c) She may or may not study late in the library.

It is important to note here that the attention of the experimenters was solely focused on option (a). Meaning, when people selected (a), they were classified as sound critical thinkers, while options (b) or (c) were bundled together as the incorrect ones. Reporting on analytical statistics for options (b) and (c) was largely absent.

However, they are quite interesting choices given a more sophisticated interpretation of the results.

| Role | Content |
|--|--|
| Conditional 1 Categorical Conclusion | If she has an essay to write, she will study late in the library. She has an essay to write She will study late in the library (MP 90%) |
| Alternative Conclusion | If she has a textbook to read, she will study late in the library. She will study late in the library (MP 94%) |
| Additional Conclusion | If the library stays open, she will study late in the library. She will study late in the library (MP 60%) |
| Conditional 1 Categorical Conclusion | If she has an essay to write, she will study late in the library. She will study late in the library She has an essay to write (AC 53%) |
| Alternative Conclusion | If she has a textbook to read, she will study late in the library. She has an essay to write (AC 16%) |
| Additional Conclusion | If the library stays open, then she will study late in the library. She has an essay to write (AC 55%) |
| Conditional 1 Categorical Conclusion | If she has an essay to write, she will study late in the library. She hasn't an essay to write. She will not study late in the library (DA 49%) |
| Alternative Conclusion | If she has a textbook to read, she will study late in the library. She will not study late in the library (DA 22%) |
| Additional Conclusion | If the library stays open, she will study late in the library. She will not study late in the library (DA 49%) |
| Conditional 1 Categorical Conclusion | If she has an essay to write, she will study late in the library. She will not study late in the library. She does not have an essay to write (MT 69%) |
| Alternative Conclusion | If she has a textbook to read, she will study late in the library. She does not have an essay to write (MT 69%) |
| Additional Conclusion | If the library stays open, then she will study late in the library. She does not have an essay to write (MT 44%) |

Table 2: Experimental Results (Dieussaert et al., 2000)

3.2.1 Interpretation and experimental procedures

There are two issues that need to be examined:

- The quality of the original experimental design.
- The explanation researchers offered for the difference between human performance and first-order Logic.

3.2.2 Confounds of the original experiment

Let us look closely at the 3 different versions of the premises given to subjects again

- (1) If she has an essay to write, she will study late in the library. (MP90%)
- (2) If the library is open, she will study late in the library. (MP60%)
- (3) If she has a textbook to read, she will study late in the library. (MP94%)

She has an essay to write

What is the main difference between the phrasing of (2) and (1),(3)? Well, (1) and (3) are strong personal reasons for going to the library. There is a specific task to be done, writing an essay or reading a textbook, for which libraries, for all intents and purposes, seem ideal. In contrast the phrasing of (2) feels general and indifferent. Is the mere fact that a library is open strong enough to go to the library? Let us rephrase this inference in another example

- (1) If she has an international conference to attend, she will go to the airport.
- (2) If the airport is open, she will go to the airport.
- (3) If she has to go on an exotic vacation, she will go to the airport.

She will go to the airport.

The point now should be quite clear. It is not a straightforward assumption that the above sentences always correspond to the same logical formulas. The choice is quite understandable in terms of experimental convenience, but it introduces certain flaws that can potentially jeopardize the validity of the conclusions that can be drawn from this experiment.

Additionally, while the original experimental design contains 3 options, (a),(b) and (c), it looks like the researchers were too focused on the answer (a) which was meant to represent to correct mathematical inference. Therefore they assumed that (b) and (c) are incorrect answers, and should be treated as such.

However, that is not necessarily the case. Option (b) is a direct negation of the first correct answer (a). It is reasonable to think that a subject choosing this answer is mentally performing an incorrect mathematical inference. But option (c) should not be equated to (b) as it is quite distinct and can potentially offer a lot of information with respect to the intent of a subject, and its perception of the cognitive task at hand.

3.2.3 Interpretation of the results

For simplicity we will take a closer look at the 1st subtask, namely the Modus Ponens (MP) one. At first, the percentage difference in the eyes of the experimenter has to be quite intriguing. Why is there such a significant statistical difference? After all, this is information that represents the same premises and therefore should be leading the subjects to draw the same conclusions. This is where the Reasoning part of this work comes into play.

Yes, if first-order Logic is assumed to be the set of logical rules that needs to be adhered to and credulous reasoning is deployed, there shouldn't be any statistical difference between the 3 versions of the same inference. But why would the experimenter ever assume these quite serious pre-conditions to hold?

Put otherwise, from an experimenters point view it is expected that when a subject is presented with a highly logical task that involves mathematical reasoning, said subject will see it for what it is: just a problem that can be computationally solved by a certain set of rules that are predetermined. For the experimenter, the meaning of the experiment is clear, but that can't be said equally about the subjects. They have to decide on the two important questions that were mentioned in the previous section which govern the way reasoning should be viewed and examined: What things are in the domain? What kind of reasoning is meant to be done about them?

In their book, Stenning and Van Lambalgen (Stenning & Van Lambalgen, 2012) align with Husserl's view, in that Logic is "simultaneously formal *and* relative to a domain", and that cognitive science needs to take into account semantics in a much more serious manner.

While the researchers were quite sensitive to the way subjects interpret sentences, they were not quite attentive to their own biases for the way the sentences were presented. Because if meaning is not given, but constructed, one has to take into account that for every cognitive task that is presented in an experiment, subjects actively struggle to impose a meaning on it. This active process has the potential to explain trains of thought that was previously deemed irrational, as perfectly rational, given a certain context.

Additionally, in the Closed-World Reasoning paradigm we assume that what is currently not known to be true is false. So there is a positive bias towards information that is true and also known to be true. Therefore the way we introduce information matters in 2 significant ways:

1. What information we chose to reveal.
2. What information we did not reveal but is relevant to the context and can alter the meaning previously revealed information.

If all information the subjects had about a character is just a sentence about having an essay to write and a library, then the information is framed in a biased way. In another context, where the character could supposedly be working in a field unrelated to writing essays, and this was the first and only essay she/he had to write maybe people would reason to different conclusions.

4 Experimental Design

It is worth noting that in the original Suppression Task experiment 24 subjects were randomly assigned to three groups. Each group received a different version of the modus ponens inference as the experimenters designed three different versions of the same arguments (simple, alternative, additional) (Byrne, 1989).

In this research, a within-subject design is going to be adopted in a form of a simple questionnaire that will be distributed online through Google Forms. This will help increase the statistical prowess of the experiment as all participants will receive the same questions depicting a mathematical inference presented in natural language. It is expected, then, that it will be easier to achieve a target subject size of about 24 so that results can be comparable between experiments.

There was a broad search for a “guideline”, or at least some point of reference, on how to properly phrase a mathematical inference in natural language, but apparently not much work has been done towards this direction. On the contrary, most of the literature concerning these topics is NLP related, meaning decoding inferences from natural language into a more mathematical system.

Therefore, a different tool will be used to create variance in the phrasing of the inferences. The dependent variable will be the framing (phrasing, context and semantics) of the inference, while the independent variable will be the kind of argumentation we are depicting, namely a Modus Ponens one.

4.1 Framing

“Framing” is a concept that refers to situations where a speaker wants to communicate a message and selects a multitude of possible variations of presenting the same information to the listener. Often, the selection of the variations is done in order to promote the speaker’s agenda on a certain issue.

For example, in an experimental study it was found that consumers view more favorably beef products labeled 75% lean, as opposed to 25% fat (Levin & Gaeth, 1988).

There is rich literature on the efficacy of framing and its effects since it was first introduced by Tversky and Kahneman in 1981 (Tversky & Kahneman, 1981). They were able to demonstrate systematic reversals of preference when the same problem is presented in different ways. In general, Sher and McKenzie argue that

Framing experiments seek to rigorously separate out the effects of relevant and irrelevant information on human judgment and choice processes ... framing effects provide a compelling reason to separate descriptive from normative models of choice. It is surely rational to treat identical problems identically, but often people do not. (Sher & McKenzie, 2011)

The description above seems to align quite well with the purposes of this research, as it aims to investigate models of human cognition by clarifying reasons that justify the normative/descriptive gaps that often appear between human performance and normatively appropriate responses.

4.2 Experimental Design

In this experiment we vary the phrasing of modus ponens mathematical inferences.

Two distinct types of framing will be used.

- A-type: a strong personal reason
- B-type: a more vague/general reason

This idea is supported by two main lines of argumentation.

First, it originates from the criticisms and observations of the original phrasing of the sentences of the Suppression Task experiment that was discussed in previous chapters.

Secondly, it reflects upon the fact that there are strong reasons to believe that pure Logic might not be an appropriate normative model for this particular cognitive task, an opinion that this research aims to prove through experimental and theoretical work.

Since the questionnaires will be distributed online, it was deemed favourable that they are shorter rather than longer so that subjects do not lose motivation or get distracted when completing them. So, each questionnaire will be comprised of 10 questions in total. 4 of them will belong to group A (strong personal framing), 4 of them to group B (vague/indifferent framing) and 2 of them will be the same control questions.

Although constructing sentences that display a certain kind of framing is quite manageable, constructing sentences with no framing is a challenge harder than it might originally appear. Intuitively, we would want sentences that display factual, pure information, and nothing else. If there is use of information that is too specific like “Water boils at 100 degrees Celsius, at sea level with normal atmospheric pressure”, we might exclude subjects that might not have the appropriate educational background. Therefore, more common everyday experiences were selected, that display general factual information that should be easily manageable for the majority of the population, regardless of their prior education.

4.2.1 Research Question

This experiment is aimed at replicating the conditions of the original Suppression Task experiment. Let us look closely to the 3 different versions of the premises given to subjects again.

- (1) If she has an essay to write, she will study late in the library. (MP90%)
- (2) If the library is open, she will study late in the library. (MP60%)
- (3) If she has a textbook to read, she will study late in the library. (MP94%)

She has an essay to write

Our running hypothesis here is that the large discrepancy between version (2) and versions (1)&(3) are influenced by the same style of framing as the one proposed in our experiment. Therefore by trying to replicate such a variance in framing and test it, we could obtain experimental evidence that supports our running hypothesis.

Specifically, our main question that needs to be answered is

Can a systematic variance in the framing of modus ponens inferences produce measurable differences in response patterns amongst individuals?

A positive answer to this question would provide evidence that could signal a potential experimental flaw in the original Suppression Task experiment. It would also support the idea that pure Logic is not an appropriate normative model for this particular task, as it does not account for the differences in framing that the proposed sentences contain.

4.3 Method

4.3.1 Participants

In total there were 44 participants (32 female) that completed the online questionnaire on a voluntary basis. The participants ranged in age from 20 to 63 years ($M = 30.3$, $SD = 9.1$ years).

They were mostly college educated (77.2% claiming to have obtained at least a Bachelor's degree and 54.5% claiming to have obtained at least a Master's degree) with a good understanding of English (86.3% claiming to have an advanced level of understanding and only 4.5% claiming to have an elementary level of understanding). The majority of the participants (72.7%) claim to have no formal education in Logic (Figure 14).

Informed consent was obtained from all participants.

4.3.2 Materials Used

A questionnaire comprised of 10 questions in total was devised and distributed online via Google Forms. The questions belong into 3 distinct categories:

- **Group A** questions display a strong personal framing. They are meant to present the modus ponens inference in the most favourable light and influence subjects towards making the correct choice.
- **Group B** questions display a vague and indifferent framing. As such, they are the antithesis of group A. They are meant to present the modus ponens inference in the worst possible light and influence subjects towards making the wrong choice.
- **Control** questions display purely factual information devoid of any kind of framing. They are used for the qualitative assessment of responses.

Specifically the distribution of questions into these 3 categories is: 4 of them will belong to group A, 4 of them to group B and 2 of them are control questions.

Each question, has 3 possible choices, to be selected by the subject.

1. Choice (a) represents the correct mathematical inference.
2. Choice (b) represents the wrong mathematical inference and it is the negation of choice (a).
3. Choice (c) represents an indifferent stance. It is neither the rejection nor the acceptance of the correct mathematical inference.

Here, for the convenience of the reader we are first presenting them bunched together in their respective groups. The order was shuffled in the actual Google Forms questionnaires that were distributed online. An example can be found in the Appendix in Figures 9-13.

Group A

Upon further suggestion of prof.Rosalie there is also the use of names in order to present the inferences in a more relatable way to the subjects.

If Susy has an essay to write, she will study late in the library. Susy has an essay to write.

- (a) She will study late in the library.
- (b) She will not study late in the library.
- (c) She may or may not study late in the library.

If Peter goes out with his friends, he will stay up late. Peter has gone out with his friends.

- (a) He will stay up late.
- (b) He will not stay up late
- (c) He may or may not stay up late.

If Sarah's mailbox is full, then the mailman has been delivering mail. Sarah's mailbox is full.

- (a) The mailman has been delivering mail.
- (b) The mailman has not been delivering mail.
- (c) The mailman may or may not have been delivering mail.

If John is late for work, he will drive his car faster on the highway. John is late for work.

- (a) He will drive his car faster on the highway
- (b) He will not drive his car faster on the highway
- (c) He may or may not drive his car faster on the highway.

Group B

Here the questions are under a neutral/vague framing.

If there is a TV news programme broadcasted in the TV, he will watch the news. There is a TV news programme broadcasted in the TV.

- (a) He will watch the news.
- (b) He will not watch the news.
- (c) He may or may not watch the news.

If the train is late, she will take the bus. The train is late.

- (a) She will take the bus.
- (b) She will not take the bus.
- (c) She may or may not take the bus.

If the airport is open, he will go to the airport. The airport is open.

- (a) He will go to the airport.
- (b) He will not go to the airport.
- (c) He may or may not go to the airport.

If there's no food left on the fridge, they will order pizza. There's no food left on the fridge.

- (a) They will order pizza.
- (b) They will not order pizza.
- (c) They may or may not order pizza.

Control Group

If the sun has been up for at least half an hour, then it is daytime.

- (a) It is daytime.
- (b) It is not daytime.
- (c) It may or may not be daytime.

If it is raining the streets will get wet. It is raining.

- (a) The streets will get wet.
- (b) The streets will not get wet.
- (c) The streets may or may not get wet.

4.3.3 Procedure

First, the participants were asked for their informed consent (Figure 12).

Then, they completed questions regarding their personal profile aimed at informing the experimenter about the subjects that complete this questionnaire (Figure 13).

Lastly they were asked to proceed and complete the main part of the experiment (10 questions, Figure 9-11).

4.3.4 Design

An one-factor within-subjects design was implemented to investigate differences in response performance between Group A and Group B questions. The order of questions was randomly shuffled.

4.3.5 Measures

We measured qualitative and quantitative measures in order to make a personal profile for the participants. Specifically gathered information on

- *Gender*(Male, Female, Other)
- *Age* (numeric response)
- *Level of Education*(Basic, Secondary, Post-Secondary, Bachelor's, Master's, PhD)
- *Fluency in English*(Elementary, Intermediate, Advanced, Proficiency)
- *Formal Education in Logic*(Yes, No)

Then, the frequency of responses in all 10 questions that were depicting mathematical inferences was measured for each participant. Specifically, for all questions the frequency of selection between choices (a), (b) and (c) was measured.

4.4 Expectations

There is the expectation that people will be more inclined to approve of the suggested inference if there is a strong personal reason to do so as the framing of the inference would suggest. If this hypothesis turns out to be correct, then there will be a statistically significant difference between the two different groups of questions similar to the one that was first observed in the Original Suppression task experiment.

On the other hand, there might not be a statistically significant difference between the two groups of proposed experimental questions. Because there are a lot of reasons that might cause such an issue beyond an obvious failure of the specific experiments it would be prudent to discuss this after some data has already been collected.

Lastly in the original experiment, researchers chose disregard any data regarding the ambiguous (c) option. This was another observation that inspired this piece of research to be more thorough, so the (c) options will be fully mentioned and analyzed as they may contain valuable insights for the data that will be collected.

5 Results

10 participants failed the control questions, therefore in this analysis we will consider 34 participants in total. Figure 15 (see Appendix) displays the data that was collected in a concentrated way, for the convenience of the reader.

As it was explained in the experimental design section, choice (a) depicts the correct Modus Ponens inference, choice (b) depicts the rejection of the inference and choice (c) signals indifference towards any particular inference.

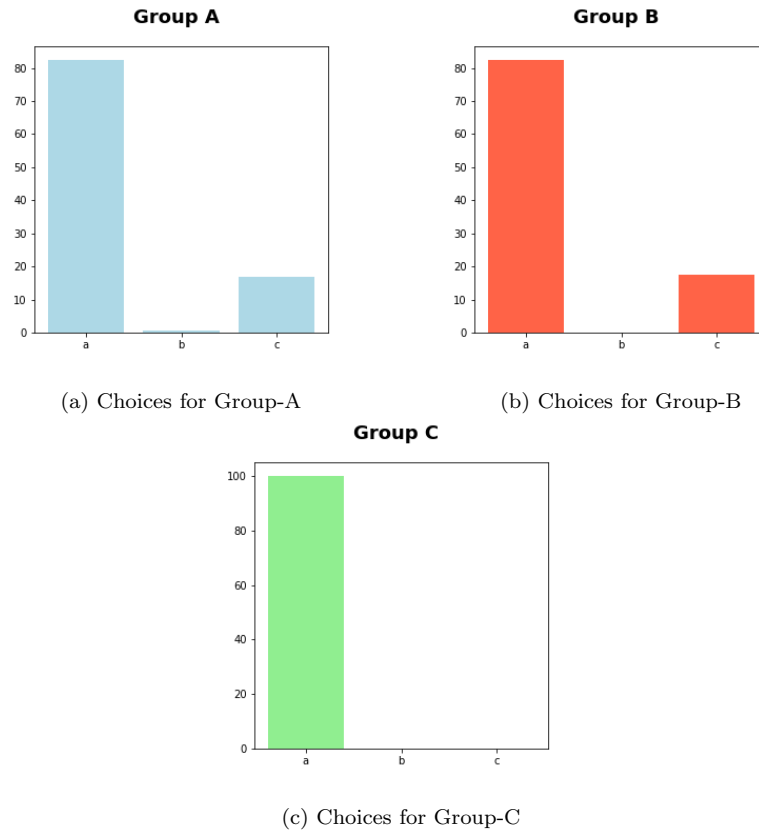


Figure 6: Choices by participants in percentages for each group of questions

5.1 Statistics

In total, participants chose the (a) response 73.86%, the (b) response 2.8% and the (c) response 23.3%.

Participants who passed the control questions chose the (a) response 82.35%, the (b) response 0.36% and the (c) response 17.3% of the time.

| | (a) | (b) | (c) |
|---------|--------|-------|--------|
| Total | 82.35% | 0.36% | 17.3% |
| Group-A | 82.35% | 0.7% | 17.03% |
| Group-B | 82.35% | 0% | 17.65% |

Table 3: Data Percentages

So sadly we can already observe that statistical results do not support our original hypothesis.

5.1.1 Inferential Statistics

Since our datasets are purely categorical we will use Chi-Squared tests to test various hypotheses. So, are there any measurable difference in responses between questions of Group-A versus Group-B ?

What seems like an obvious rejection of this hypothesis just by observing a summary of the datasets that were collected can be also verified by large p-values and an overlap of the error bars in our barplot (Figure 7).

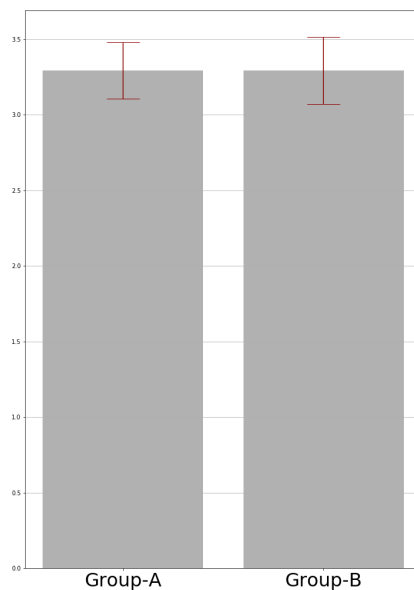


Figure 7: Barplot of the Standardized Error of the Mean between Group-A and Group-B

Specifically for our dataset we calculate
 $X^2(2, N = 34) = 1.021, p = 0.6$

5.2 Discussion on Experimental Results

Here we will consider some feedback that was attained from participants, whenever that was possible, and how it relates to our experimental results.

5.2.1 Feedback from Participants

One important difference with the original experimental conditions, is that participants in the original experiment completed the questionnaires in person. Although this is not reported in the write-up of the results, there is the assumption that in person contact might eliminate misunderstandings about the experimental procedure. Through written feedback that was obtained, it was revealed that sometimes subjects were a bit confused on what the “right answer” is supposed to be on such an experiment. This led to some unconventional thinking like:

“If the sun is up for at least 30 minutes, it should be daytime. But what if, hypothetically, were we to be located somewhere close to the North Pole? That information alone would not be enough.” (participant 4)

In the experimental design, as it was mentioned in the relevant section, control questions were meant to represent factual information without any kind of framing in them. However, in practise, it seems that this approach led a lot of participants to be highly sceptical about them.

But high amounts of scepticism appeared through out the questionnaire and they were not exclusive to the control questions. This is another example:

“If John is late for work, he will drive his car faster on the highway. John is late for work. But he might not want to drive faster because of fears from getting penalized with a speeding ticket. Or he can not drive faster anyway because he is stuck in traffic.” (participant 5)

How can we explain then, from a theoretical point of view then the high amounts of scepticism that were observed in this experiment?

Maybe we have to consider the broader context in which the experiment was pushed, which was social media. In this study (Fletcher & Nielsen, 2019), researchers found that although the majority of people may not understand how the information that reaches them is being filtered, they do not uncritically accept it. It seems that people are highly sceptical about all content that they read on social media websites, regardless of who is the publisher.

So, although an online questionnaire is not a news item, when posted in a social media group where most posts are news or advertisements, it is reasonable to expect that it might get treated the same way. This holds true for participants that do not belong in the immediate social circle of the experimenter. Then, we take into account the fact that the verbal feedback that was obtained was from people that did belong in the immediate social circle of the experimenter. Those people, although a smaller subset of the total sample, were still

worried about making the "correct choice". However, when they completed the experiment, and they received some explanation about the design and the goals of the experiment, they seemed to be approving of the project.

Even though the data collected do not validate the original hypothesis of the experiment, these observations underline the importance of context whenever we have to deal with human cognition in real life, and even in an experimental setting like this. They also validate the thesis of reasoning *to* and *from* an interpretation (Stenning & Van Lambalgen, 2012).

6 General Discussion

Since this kind of research is highly experimental, it is prudent to address some valid points a potential critic might raise, as well as investigate additional explanatory propositions.

6.1 Criticisms

Looking back at the results of this experiment, we will address the following criticisms:

1. Criticisms on the execution of the experiment.
2. Criticisms on the experimental design

There is little to argue about the execution of the experiment. For all intents and purposes, all standards about conducting online surveys were followed. The only meaningful factor that might have played a crucial role was that the experiment took place online, instead of offline, like the original Suppression Task experiment. This point was addressed previously and given the circumstances in which this study is taking place, no alternative course of action could have been taken.

On the other hand, there are some matters to discuss about the design and the general inspiration for running this type of an experiment.

Let us address first one hypothesis under which the experiment subliminally operates, which we will call the “No framing” hypothesis. This is the assumption that there can be information that is transmitted and presented without any kind of specific frame attached to it.

Obviously, if the reader rejects the possibility that neutral framing exists, there is no counter arguments that can be offered. The experimental design becomes, then, problematic in principle and so is part of this research.

However, accepting the possibility of neutral framing will lead a potential critic to a path that is, if nothing else, extremely intriguing. There is a lot of relevant literature, especially in the field of experimental economics, that acknowledges the fact that information can be presented in a neutral way, although no formal proofs can be found. In this study, researchers conducted a context-less auction for a meaningless good (Dürsch & Müller, 2017). They found that

“ ...overly neutral instructions, which lack any contextual clues, can lead to strange behavior. In a contextless second price auction for a meaningless good, a majority of subjects enter positive bids—a case of cognitive experimenter demand effect. Subjects bid positive amounts because this is what they think they are tasked with in the experiment. Adding a second auction that has a context drastically reduces the positive bids in the meaningless first auction by reducing the cognitive experimenter demand effect. ” (Dürsch & Müller, 2017)

So, then it can be argued that neutral framing is discouraging the subjects from accepting the correct mathematical inference, something that is opposite to our original assumptions. This can be a valid point, as it is not certain that displaying bare factual information would be highly favorable to making the correct inference.

It is also reasonable to suggest that there might be objections to the use of framing in general, with respect to this experimental design. That is to say that creating variation in mathematical inferences through framing is not advisable. However, the purpose of this research is to investigate the effects that context and semantics have on reasoning and therefore on cognitive models. As it was mentioned on the experimental design section, there are no universal linguistic guidelines on how to transfer an inference from a mathematical symbolic language to natural language. Therefore, while this is another valid criticism, it would have to come with an alternative suggestion on how to create variance in inferences.

Finally, another observation is that the questions that were used are relatively easy to answer. This, combined with the fact that (a) was always the correct answer might have introduced a hidden confound in the experiment as people were always incentivized to always click option (a). So an alternative experimental design could be proposed were in some questions option (b) is the correct answer. This could be achieved through the use of negation. For example, this original question

If Susy has an essay to write, she will study late in the library. Susy has an essay to write.

- (a) She will study late in the library.
- (b) She will not study late in the library.
- (c) She may or may not study late in the library.

can be transformed so that now (b) is the correct answer like so

If Susy has to work at the restaurant, she will not study late in the library. Susy has to work at the restaurant.

- (a) She will study late in the library.
- (b) She will not study late in the library.
- (c) She may or may not study late in the library.

6.2 Exploratory Analysis

In order to adhere to rigorous scientific standards, the experimental design along with the research hypothesis, should remain the same throughout the official analysis of the results. Since that has already taken place, in this section we reflect on what was done and investigate additional ideas outside the scope of the original experiment.

One of the main criticisms we addressed in the previous section had to do with the design of the control questions and the “No Framing” hypothesis. So, what would happen if we treated then every question of group B and C the same? Meaning we assume the stance that neutral framing is impossible and our control questions really fall under the category of a vague/indifferent framing.

Under this constraint, we would have 4 Group-A versus 6 Group-B (4 original plus the 2 Control questions) from the data that was collected. For convenience we will refer to Group-A questions as (A1, A2, A3, A4), Group-B questions as (B1, B2, B3, B4, C1, C2). To surpass this class imbalance we chose to do random sampling as follows:

We count unique combinations without repetition of elements of the set (B1, B2, B3, B4, C1, C2) in groups of 4. This will produce 15 sets that will look like this

$$\{ (B1, B2, B3, B4), (B1, B2, B3, C1), (B1, B2, B3, C2), \dots, (B3, B4, C1, C2) \}$$

We then aggregate responses across the different subsets of datasets that would correspond to each unique combination of Group-B questions and average them out. The results are displayed in Table 4.

| | (a) | (b) | (c) |
|---------------|--------|-------|--------|
| Total | 76.33% | 2.84% | 20.83% |
| Total(Type-A) | 76.13% | 2.27% | 21.59% |
| Total(Type-B) | 76.15% | 3.4% | 20.01% |

Table 4: Aggregated Data Percentages

Percentages are slightly better compared to the ones in Table 2, as, for example, more people reject the correct inference in Group-B questions(3.4%) compared to Group-A questions(2.27%). However, they are still far from being able to support with any measurable degree of confidence our original hypothesis.

6.3 Relation to The Most Common Explanations of Irrationality

In previous chapters we examined some of the most common explanations of irrationality. Those were (Stanovich, 1999):

- Systematic irrationality in the intentional-level psychology
- Algorithmic level limitations
- Performance errors
- Incorrect norm application
- Alternative problem construal

Here we will relate these explanations to the specific results of this particular experiment, in an attempt to gain further insight on our observed results.

Admittedly, there were few instances where the subject’s behavior diverged from our original expectations. This divergence was expressed mainly by high amounts of skepticism regarding the experimental task. We have mentioned before, that rationality in experimental tasks is relative to the normative model an experimenter chooses. For the purposes of this experiment, we deemed choice (a) as rational, as it depicts the correct mathematical inference according to first-order Logic. As it can be seen in the tables of Figure 8, the majority of participants, more often than not, made the correct choice.

| | (a) | (b) | (c) |
|---------|--------|-------|--------|
| Total | 82.35% | 0.36% | 17.3% |
| Group-A | 82.35% | 0.7% | 17.03% |
| Group-B | 82.35% | 0% | 17.65% |

| | (a) | (b) | (c) |
|---------|--------|-------|--------|
| Total | 76.33% | 2.84% | 20.83% |
| Group-A | 76.13% | 2.27% | 21.59% |
| Group-B | 76.15% | 3.4% | 20.01% |

(a) Original Dataset

(b) Resampled Dataset

Figure 8: Comparison of Original and Resampling Results

Due to this fact, it would be hard to ascribe any form of systematic irrationality to the participants. Then, the task was not at all taxing from a computational cost perspective. So, again, it is hard to identify any algorithmic level limitations participants would face in this experiment.

Going down our list of potential explanations for the normative/descriptive gap, we continue with performance errors. The reader should be reminded here, of the main characteristics of performance errors, as they were unidentified in Chapter 2 of this study. Those characteristics were, *odness*, *independence* and *randomness* (Wang, 2008). Performance errors might have happened, however as it is usually the case with performance errors, it is extremely difficult to identify them. Our best guess would be that, since the task was relatively simple and not cognitively taxing, we can assume performance errors were few

and far between amongst subjects and have little to no explanatory capability for the observed results.

On the contrary, alternative problem construal and incorrect norm application, are two explanations that can have a lot of merit, and in our case they go hand in hand. We base this assumption on the high amounts of skepticism regarding the experimental task subjects often displayed. In fact, highly skeptical subjects would often view option (c) (indifference towards the modus ponens inference) as the correct answer. For these subjects then, sentences were not perceived purely as displaying a modus ponens mathematical inference but potentially a more general real life situation in which context matters and outcomes are uncertain. This perception would obviously lead them to not apply pure first-order logic in order to come to the optimal solution.

Of course, it is debatable what other method the subjects would then deploy in order to answer the questions, but this is one of the problematic issues that this thesis often comes with, as we have stated previously in Chapter 2 where we explored the concepts of alternative problem construal and incorrect norm application in depth. But, despite the fuzziness of how exactly those two concepts would work in this particular case, we at least have enough evidence to identify them, and therefore ascribe to them some of the cause for observed normative/descriptive gap.

6.4 Closing Remarks

This study focuses on exposing the role of context and semantics in reasoning, especially in syllogistic tasks that involve mathematical inferences. Although our pure experimental results did not validate our original hypothesis, some of the exploratory analysis, along with written feedback that was obtained by some participants, do offer inspiration for further experimentation. For example if instead of constructing sentences that vary a kind of framing, what if we constructed sentences with framing and sentences without. Would the results change? There is evidence in the exploratory analysis done above that could support this case.

Then, another idea could be that the use of mathematical inferences depicted by sentences in natural language is not the optimal experimental design to explore the role of context and semantics in reasoning. We can look at instances of experimental work done in economics like this paper (Dürsch & Müller, 2017), where researchers made up context-less auctions. An auction or a game can offer a more robust way of testing a certain kind of framing that can surpass the limitations of natural language.

As far as AI is concerned, context and semantics are notoriously hard to model. Critics of the strong version of AI argue that computer systems might be unable to model such concepts in principle. Therefore any advancement that helps build and inform more accurate human cognitive models should also be a step towards making the vision of *strong* AI a reality. This work hopes to make a contribution, however small it may be, towards this direction.

Appendices

The image shows a screenshot of a Google Form titled "Experiment regarding mathematical inferences". The form has a blue header with a decorative background of binary code and circuitry. Below the title, there is a red asterisk and the word "Required". The form is divided into three sections, each with a blue header "Section of questions". Each section contains a conditional statement followed by three radio button options. The first section's statement is "If Susy has an essay to write, she will study late in the library. Susy has an essay to write. *". The second section's statement is "If Peter goes out with his friends, he will stay up late. Peter has gone out with his friends. *". The third section's statement is "If there is a TV news programme broadcasted in the TV, he will watch the news. There is a TV news programme broadcasted in the TV. *". Each section has three radio button options: "She will study late in the library", "She will not study late in the library.", "She may or may not study late in the library" for the first; "He will stay up late", "He will not stay up late", "He may or may not stay up late" for the second; and "He will watch the news.", "He will not watch the news", "He may or may not watch the news" for the third. A small edit icon is visible in the bottom right corner of the form.

Experiment regarding mathematical inferences

* Required

Section of questions

If Susy has an essay to write, she will study late in the library. Susy has an essay to write. *

She will study late in the library

She will not study late in the library.

She may or may not study late in the library

If Peter goes out with his friends, he will stay up late. Peter has gone out with his friends. *

He will stay up late

He will not stay up late

He may or may not stay up late

If there is a TV news programme broadcasted in the TV, he will watch the news. There is a TV news programme broadcasted in the TV. *

He will watch the news.

He will not watch the news

He may or may not watch the news

Figure 9: Questionnaire in Google Forms as seen by the participants

The image shows a Google Forms questionnaire with five questions, each with three radio button options. The questions are conditional logic questions. The first question is: "If the sun has been up for at least 30 minutes, then it is daytime. The sun has been up for at least 30 minutes. *". The options are: "It is daytime", "It is not daytime", and "It may or may not be daytime". The second question is: "If the train is late, she will take the bus. The train is late *". The options are: "She will take the bus", "She will not take the bus", and "She may or may not take the bus". The third question is: "If Sarah's mailbox is full, then the mailman has been delivering mail. Sarah's mailbox is full *". The options are: "The mailman has been delivering mail", "The mailman has not been delivering mail", and "The mailman may or may not be delivering mail". The fourth question is: "If the airport is open, he will go to the airport. The airport is open. *". The options are: "He will go to the airport", "He will not go to the airport", and "He may or may not go to the airport". The fifth question is: "If it is raining, the streets get wet. It is raining *". The options are: "The streets will get wet" and "The streets will not get wet". There is a small chat icon in the bottom left and a pencil icon in the bottom right of the form area.

If the sun has been up for at least 30 minutes, then it is daytime. The sun has been up for at least 30 minutes. *

It is daytime

It is not daytime

It may or may not be daytime

If the train is late, she will take the bus. The train is late *

She will take the bus

She will not take the bus

She may or may not take the bus

If Sarah's mailbox is full, then the mailman has been delivering mail. Sarah's mailbox is full *

The mailman has been delivering mail

The mailman has not been delivering mail

The mailman may or may not be delivering mail

If the airport is open, he will go to the airport. The airport is open. *

He will go to the airport

He will not go to the airport

He may or may not go to the airport

If it is raining, the streets get wet. It is raining *

The streets will get wet

The streets will not get wet

Figure 10: Questionnaire in Google Forms as seen by the participants

If the airport is open, he will go to the airport. The airport is open. *

He will go to the airport

He will not go to the airport

He may or may not go to the airport

If it is raining, the streets get wet. It is raining *

The streets will get wet

The streets will not get wet

The streets may or may not get wet

If John is late for work, he will drive his car faster on the highway. John is late for work. *

He will drive his car faster on the highway

He will not drive his car faster on the highway

He may or may not drive his car faster on the highway

If there's no food left on the fridge, they will order pizza. There's no food left on the fridge. *

They will order pizza

They will not order pizza

They may or may not order pizza

[Back](#) [Submit](#)

Never submit passwords through Google Forms.

This form was created inside of Universiteit Utrecht Studenten. [Report Abuse](#)




 

Figure 11: Questionnaire in Google Forms as seen by the participants



Experiment regarding mathematical inferences

Thank you for taking part in this experiment.
In this experiment, you will be asked to answer questions regarding mathematical inferences. You will read statements and asses whether you agree with them or not.

The experiment should take you at most 3 minutes.

Your participation is voluntary, and you are free to stop taking part at any point. Your data will be processed anonymously, and will never be traced back to you individually. Your data will be used for research purposes and communication of research. Data might be shared for such purposes, but will always be anonymous.

If you have any further questions, you can contact s.kantiloros@students.uu.nl

* Required

I have read the above information and agree with it. By ticking this box I consent to take part in this experiment *

Yes

No

[Next](#)

Never submit passwords through Google Forms.

This form was created inside of Universiteit Utrecht Studenten. [Report Abuse](#)

Google Forms

Figure 12: Informed consent part of the questionnaire in Google Forms as seen by the participants

Personal profile

What is your gender ?

- Male
- Female
- I do not want to specify
- Other: _____

What is your age?

Your answer _____

What is your level of education?

- Basic education
- Secondary education
- Post-secondary education
- Bachelor's degree
- Master's degree
- Doctoral degree
- Other: _____

What is your level of fluency in English?

- Elementary Level
- Intermediate Level
- Advanced Level
- Proficiency Level / Native speaker

Your answer _____

What is your level of education?

- Basic education
- Secondary education
- Post-secondary education
- Bachelor's degree
- Master's degree
- Doctoral degree
- Other: _____

What is your level of fluency in English?

- Elementary Level
- Intermediate Level
- Advanced Level
- Proficiency Level / Native speaker

Have you ever taken any classes in logic?

- Yes
- No

[Back](#) [Next](#)

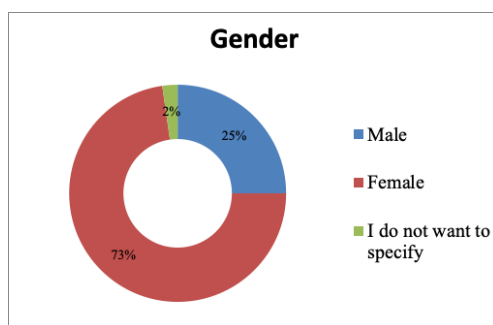
Never submit passwords through Google Forms.

This form was created inside of Universiteit Utrecht Studenten. [Report Abuse](#)

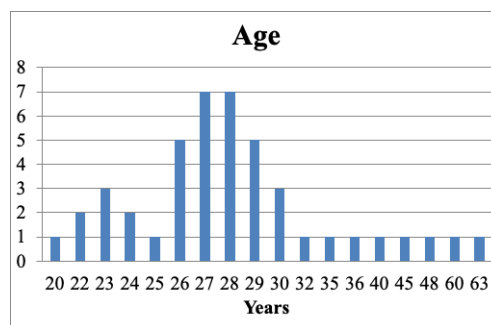
Google Forms

Figure 13: Personal profile questions in Google Forms as seen by the participants

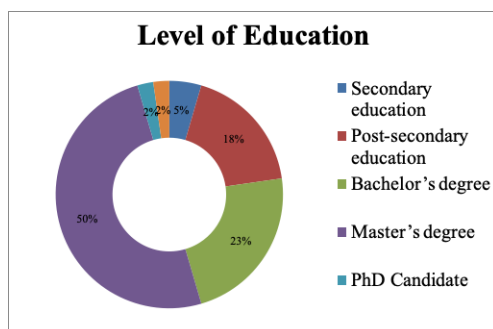
The Makeup of Participants



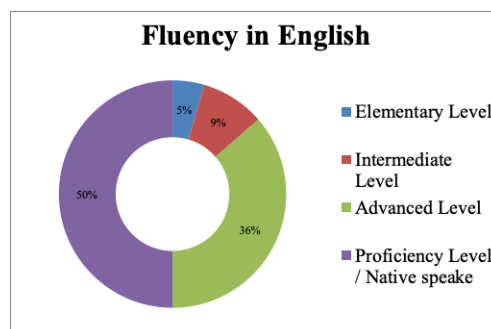
(a) Gender



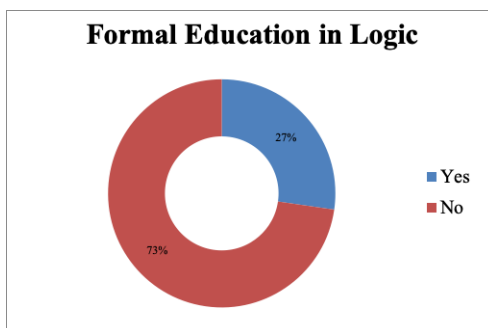
(b) Age



(c) Level of Education



(d) Fluency in English



(e) Formal Education in Logic

Figure 14: Makeup of Participants

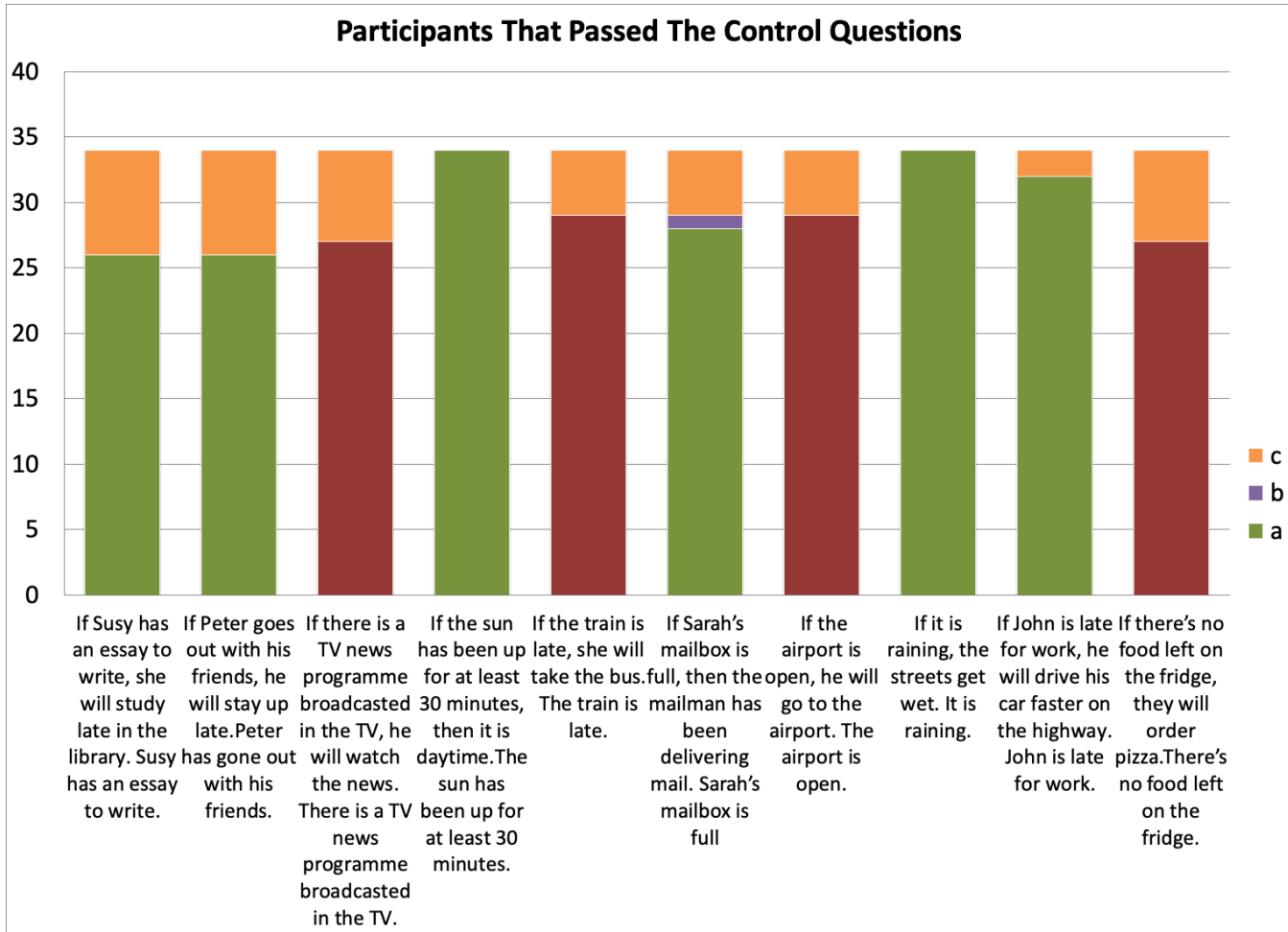


Figure 15: Responses from Participants

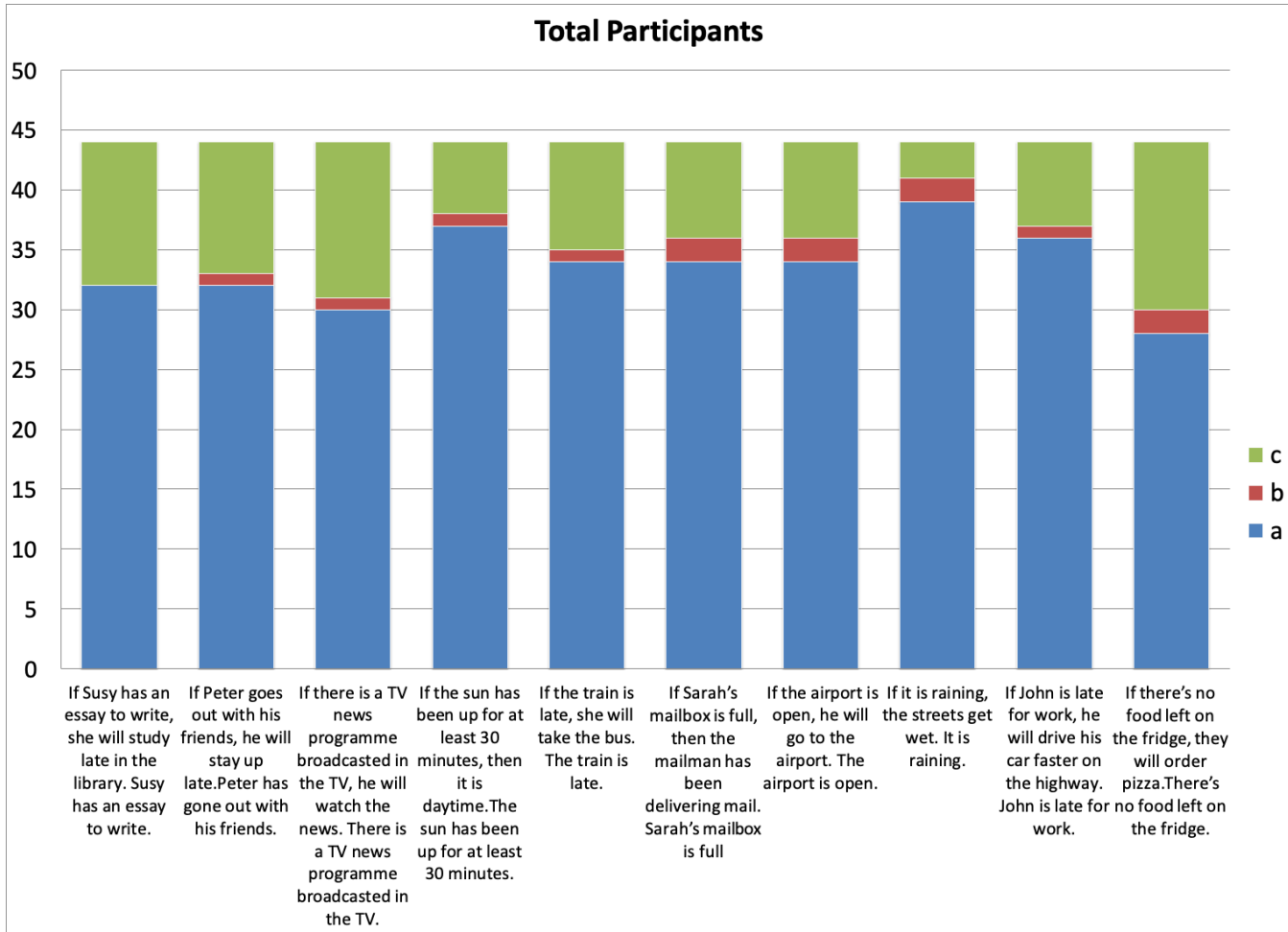


Figure 16: All participants

References

- Baron, J. (2012). The point of normative models in judgment and decision making. *Frontiers in Psychology, 3*, 577. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2012.00577> doi: 10.3389/fpsyg.2012.00577
- Bell, B. G., Gardner, M. K., & Woltz, D. J. (1997). Individual differences in undetected errors in skilled cognitive performance. *Learning and Individual Differences, 9*(1), 43–61.
- Byrne, R. M. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*(1), 61–83.
- Dieussaert, K., Schaeken, W., Schroyens, W., & d’Ydewalle, G. (2000). Strategies during complex conditional inferences. *Thinking & Reasoning, 6*(2), 125–160.
- Dürsch, P., & Müller, J. (2017). Bidding for nothing? the pitfalls of overly neutral framing. *Applied Economics Letters, 24*(13), 932–935.
- Elqayam, S., & Evans, J. S. B. (2011). Subtracting” ought” from” is”: Descriptivism versus normativism in the study of human thinking. *Behavioral and Brain Sciences, 34*(5), 233.
- Elqayam, S., & Over, D. E. (2016). From is to ought: The place of normative models in the study of human thought. *Frontiers in psychology, 7*, 628.
- Fletcher, R., & Nielsen, R. K. (2019). Generalised scepticism: how people navigate news on social media. *Information, Communication & Society, 22*(12), 1751–1769.
- Haugeland, J. (1985). *Artificial intelligence: the very idea*. Cambridge, MA: MIT Press.
- Hertwig, R., & Gigerenzer, G. (1999). The ‘conjunction fallacy’ revisited: How intelligent inferences look like reasoning errors. *Journal of behavioral decision making, 12*(4), 275–305.
- Levin, I. P., & Gaeth, G. J. (1988). How consumers are affected by the framing of attribute information before and after consuming the product. *Journal of consumer research, 15*(3), 374–378.
- Messer, W. S., & Griggs, R. A. (1993). Another look at linda. *Bulletin of the Psychonomic Society, 31*(3), 193–196.
- O’Hare, L. (2004). Measuring critical thinking skills and dispositions in undergraduate students..
- Rips, L., & Conrad, F. (1983). Individual differences in deduction. *Cognition and Brain Theory, 6*, 259–285.
- Russell, S., & Norvig, P. (2002). *Artificial intelligence: a modern approach*.
- Sher, S., & McKenzie, C. R. (2011). Levels of information: A framing hierarchy. *Perspectives on framing, 35*.
- Stanovich, K. E. (1999). *Who is rational?: Studies of individual differences in reasoning*. Psychology Press.
- Stanovich, K. E., & West, R. F. (1998). Individual differences in rational thought. *Journal of experimental psychology: general, 127*(2), 161.

- Stein, E. (1996). *Without good reason: The rationality debate in philosophy and cognitive science*. Clarendon Press.
- Stenning, K., & Van Lambalgen, M. (2012). *Human reasoning and cognitive science*. MIT Press.
- Stevenson, A. (2010). *Oxford dictionary of english*. Oxford University Press. Retrieved from <https://www.oxfordreference.com/view/10.1093/acref/9780199571123.001.0001/acref-9780199571123> doi: 10.1093/acref/9780199571123.001.0001
- Stupple, E. J. N., & Ball, L. J. (2014). The intersection between descriptivism and meliorism in reasoning research: further proposals in support of ‘soft normativism’. *Frontiers in Psychology*, 5, 1269. Retrieved from <https://www.frontiersin.org/article/10.3389/fpsyg.2014.01269> doi: 10.3389/fpsyg.2014.01269
- Tversky, A., & Kahneman, D. (1981). The framing of decisions and the psychology of choice. *science*, 211(4481), 453–458.
- Tversky, A., & Kahneman, D. (1983). Extensional versus intuitive reasoning: The conjunction fallacy in probability judgment. *Psychological review*, 90(4), 293.
- Wang, Y. (2008). On cognitive properties of human factors and error models in engineering and socialization. *International Journal of Cognitive Informatics and Natural Intelligence (IJCINI)*, 2(4), 70–84.
- Wason, P. C. (1966). Reasoning en, b. foss (comp.). *New horizons in psychology*, 135–151.
- Wetherick, N. (1995). Reasoning and rationality: A critique of some experimental paradigms. *Theory & Psychology*, 5(3), 429–448.