Eötvös Loránd University

Faculty of Science

Department of Computer Science

Logikatanítás a középiskolában

Tanulás egy bizonyítássegítő programmal

To Learn Logic Better, Teach the Computer First

Learning with an Automated Theorem Prover and Human Recognition of Inference Rules in High School

Author:

Ádám Armand Ficker

Angol nyelv és kultúra – Matematika tanár osztatlan tanári mesterszak

Internal supervisor:

Péter Sziklai

Full professor, ELTE

External supervisor:

Zoltán Gábor Molnár

Adjunct professor, BME

Budapest, 2021

Contents

Chapter 1

Introduction

1.1 Formal Reasoning and High School Mathematics

We are in a difficult position when we examine the part of high school mathematics curriculum that deals with formal reasoning. This may be surprising, since according to Aristotle, mathematics is primarily an argumentative, proof based discipline. [1] Nonetheless, reasoning, as a kind of argumentative tool, which, in fact, should be considered to be the the most fundamental device to establish and verify the truth of mathematical theorems, is strangely enough mostly discussed in Hungarian Literature classes rather than in Mathematics classes. Of course, all the mathematics course books devote time to prove the most important statements, even if not all of them. And indeed, logic itself is discussed, at least to some degree, in the twelfth grade material, over which, however, the impending school-leaving exam casts an ominous shadow. Additionally, the ninth grade mathematics curriculum has always featured a chapter titled "Methods of Proof", since the time of Tamás Varga. [2] But these excerpts serve only a role of curiosities, rather than a topic to be mastered and explored in detail. Clearly, the traditions of teaching mathematics in Hungary, which has culminated in the heuristics of György Pólya, presupposes not only a practical knowledge of proving statements, but even the knowledge of what conjectures lead to a correct proof. [3] However, we cannot say that every Hungarian high school keeps this vision of mathematics in their crosshairs or that they are actively working towards creating a school system and a learning environment that the followers of György Pólya and Tamás Varga had previously dreamed of. While in the so-called "elite schools", the mathematizer ability of students is maintained at the same level, which is, in fact, outstandingly high as well, but according to PISA results, in other schools of the country this ability shows a declining trend and even more worryingly Hungarian schools cannot seem to address the differences generated by the socio-cultural background of students. [4] The situation in general may be similar, or perhaps even worse, than what was reported in English schools in the 1990s: students perceive mathematics as a kind of empirical science rather than an argumentative one. [5] Thus, teaching correct reasoning and systematic thinking is a much more essential task for the public education of today, than it has ever been. The increasingly better automated theorem prover and proof manager software packages are great for achieving both goals, since they appear to be mathematically and intuitively applicable in mathematics classes, and they have all the benefits of a rigorous programming language.

1.2 Why Coq?

The main inspiration for this thesis came from two papers: one was "Interactive Teaching of Programming Language Theory with a Proof Assistant" by Bereczky Péter, Donkó István, Horpácsi Dániel, Kaposi Ambrus and Németh Dávid János and the other one was "Learning how to Prove: From the Coq Proof Assistant to Textbook Style" by Sebastian Böhne and Christoph Kreitz. [6][7] Both of these papers focus on the struggle that university students have with formalisation in mathematics, and more precisely, in Computer Science education at university level. I believe that a large part of this difficulty stems from the lack of logic and formal reasoning taught in high schools, and in our case, more specifically, Hungarian high schools. Both of these papers use Coq as a tool to enhance these skills of the examined students, and I will attempt to do the same in my thesis.

The Coq proof assistant is a formal proof management system, first released in 1989. It provides an environment to write mathematical definitions, theorems and proofs. It also features a Proof mode, which can be used to prove theorems with the use of Coq's built-in language, which is an extended version of the Calculus of Inductive Constructions. This mode is the main reason I chose Coq as a means to teach some of the fundamental elements of Classical and Intuitionistic Logic to the high school students who took part in my experiment. The proof mode is a great way to visualize how mathematical formalisation works and the way mathematical theorems are proven step-by-step. Both formalisation and proofs are worryingly neglected in the national curriculum of mathematics in Hungary, and high school teachers are highly disincentivized to allocate any time to teach them, since it is not a requirement for the lower level matura exam, and even at the higher level exam, it is a very marginal topic, where only a really superficial level of understanding is expected from the examinees. Obviously, the level of abstraction Coq requires may prove to be a difficult challenge to tackle for many students but the interactive environment Coq provides and the syntax's similarities to mathematical notations make it more than worthwhile to spend time understanding it.

1.3 The classical and intuitionistic deductive reasoning

The focus of this thesis is limited to propositional logic and it considers deduction as a framework for natural derivation as a cognitive-mathematical model. It is cognitive if we assume that the brain can somehow organize reasoning in this way, and it is mathematical if we use a well-established conceptual system in the field of mathematical logic. There is a stance in the cognitive psychological literature on deduction, but one that is surprisingly underused, which states that the structure of logic, called "natural deduction", closely resembles the mental model used by the human brain. [8] [9]

The language of propositional logic is given in Backus–Naur style as commonly used in the study of programming languages.

$$\mathcal{L} ::= A, B, C, \dots, \bot \mid A \land B \mid A \lor B \mid A \to B$$

$\mathcal{L}$ is said to be the set of all propositions and Greek capital letters are used to denote sets of propositions (premises) $(\Gamma, \Delta, \dots)$. Sets of premises are sometimes called contexts and they could also be defined in the Backus-Naur style.

$$\mathcal{C} ::= \quad \mid \Gamma, A$$

where A is any proposition and "," is the context expanding operator.

The binary provability relation $\Gamma \vdash A$ is defined recursively according to the following table.

$$\frac{}{\Gamma, A \vdash A} \qquad \frac{\Gamma \vdash \bot}{\Gamma \vdash A}$$

$$\frac{\Gamma, A \vdash B}{\Gamma \vdash A \to B} \qquad \frac{\Gamma \vdash A \to B \qquad \Gamma \vdash A}{\Gamma \vdash B}$$

$$\frac{\Gamma \vdash A \quad \Gamma \vdash B}{\Gamma \vdash A \wedge B} \qquad \frac{\Gamma \vdash A_1 \wedge A_2}{\Gamma \vdash A_i}$$

$$\frac{\Gamma \vdash A_i}{\Gamma \vdash A_1 \vee A_2} \qquad \frac{\Gamma \vdash A \vee B \quad \Gamma, A \vdash C \quad \Gamma, B \vdash C}{\Gamma \vdash C}$$

These are the rules of natural deduction and what we see is the so-called intuitionistic propositional logic **NI**, given in the style of natural deduction. Negation can also be expressed by absurdity:

$$\neg A = A \to \bot$$

However, hitherto, the logic is not classical. It becomes classical, if we assume the law of double negation.

$$\boxed{\frac{\Gamma \vdash \neg\neg A}{\Gamma \vdash A}}$$

As we follow the logic of Coq in preparing the experimental group, it is unnecessary to better present this logic, and the specific procedures will be described later. Besides, we do not consider it appropriate to teach intuitionistical logic without Coq, merely using the whiteboard in the conventional. In the experiment, we chose a much more interactive method by working with an proof management software.

1.4 Inductive and deductive reasoning in cognitive psychology

A well-established experimental paradigm in the cognitive literature for the separation of induction and deduction is the signal detection theory, initiated by Rips and analysed by Lee's generative models. This will be used to separate the two types of deductive logic. The basics of the distinction between induction and deduction states that induction deals with plausible inferences, whereas deduction focuses on more controlled, logical conclusions. An argument can be analysed both from an inductive and a deductive standpoint. From a deductive point of view, an argument can be valid or invalid, based on a set of well-defined logical criteria. According to Skyrms [10], the definition of deductive validity is the following:
"An argument is deductively valid if and only if it is impossible that its conclusion is false while its premises are true."
On the other hand, an argument can be evaluated from an inductive perspective too, it can either be inductively strong or inductively weak. Skyrms defines it as:
"An argument is inductively strong if and only if it is improbable that its conclusion is false while its premises are true, and it is not deductively valid. The degree of inductive strength depends on how improbable it is that the conclusion is false while the premises are true."
These definitions clearly distinguish two types of arguments but it is still an unanswered question whether there are two kinds of reasoning as well. According to Heit and Rotello [11], there are two main stances regarding this question. Some researchers claim that there is a common set of reasoning processes used for evaluating both inductive and deductive arguments. They also state that people essentially reason in a non-deductive nature.
Conversely, many researchers argue that there is a distinction between two separate kinds of reasoning, one that is fast and largely context-based and another that is slower, more meticulous, more analytic and is based on a clear set of rules. These two types of reasoning do not necessarily correspond directly to inductive and deductive arguments but deductive arguments would likely rely more on the latter, whereas inductive ones would likely rely more on the former type of reasoning. The

former theory is backed by a number of experiments and the latter is supported by some neuropsychological evidence, both detailed by Heit and Rotello, thus neither approach can be entirely rejected.

Unlike these proposals, which attempt to explain certain phenomena regarding reasoning, Rips [12] focuses on distinguishing induction and deduction. He claims that there is a single scale for evaluating arguments, this account is referred to as the Criterion-shift account. This scale shows the strength of an argument, with two major dividing lines on it, the first separates inductively weak and inductively strong arguments, and the second one separates deductively valid and invalid ones. This account coincides with the definitions laid out by Skyrms for the most part, with one difference, Skyrms' definition of deduction requires a clear distinction between induction and deduction, rather than labeling deduction as just a stricter type of induction, which is the stance that Rips' approach supports. As Skyrms states:

"An argument is inductively strong...and it is not deductively valid."

Rips claims that there might be arguments that are inductively strong but not strong enough to be judged deductively valid, but if a statement is deductively valid then it is necessarily inductively strong as well, and this is the main difference between the definitions of Skyrms and the views of Rips, otherwise the Skyrms' definitions are applicable here too.

One advantage of Rips' approach is that it makes a number of testable predictions. One of these predictions is that the relative ordering of two arguments should be the same regardless of whether they are judged based on their deductive correctness or inductive strength. Hence if argument A is more likely to be judged deductively correct than argument B, then it should also be more likely to be judged inductively strong as well.

Heit and Rotello, along with Rips carried out experiments, both pointing toward the validity of two-process reasoning, and one of the main aims of my thesis is to find further results in this area, either reinforcing or questioning what they found. My research is heavily based on their methodology, with some of the shortcomings of Rips' research, detailed by Heit and Rotello, being attended to. Lee[13] also states that even though Heit and Rotello's methods are used and accepted by the majority of researchers, he highlights some weaknesses of theirs regarding their statistical analysis, mainly their disregard of individual differences, whether it was intentional or not. With these issues kept in mind, I adhered to Lee's remarks and applied his insight to use Bayesian inference to account for individual differences as well, further detailed in the 'Theoretical and Technical Backgrounds' chapter.

Further motivation to this thesis was that Heit and Rotello claim that there is a striking parallel between two-process accounts of reasoning and two-process accounts of recognition memory, the latter having already been heavily researched, providing a good case against a single process account of reasoning too. And with two-process accounts of recognition being so deeply and successfully researched, it suggests that conducting further research in this area has a high potential of producing valuable results.

Chapter 2

Theoretical and Technical Backgrounds

Lance Rips' innovation in recognizing valid inferences was that each inference functions as a unique, although complex, signal; and the well-known psychological theory of signal detection prevails in the phenomenon of logical perception. Heit and Rotello modified Rips' experiment so that the inferences were abstract. Later, I also asked high-school students such questions (see Appendix A). As an example, consider the following argument, that I will call Classical Modus Tollens (CMT).

If Paul is not A, then Paul in not B.
Paul is B.

Paul is A.

We can then ask whether this form of argument is valid, both from a deductive and an inductive standpoint. When asked if it is deductively valid, according to the experimental paradigm, we ask the subjects whether the inference is *necessary*. If it is recognized as necessary by the subjects, the argument is perceived as signal, if not, it is perceived as noise. In the case of induction, this question concerns whether the conclusion is *possible,* provided that the premises are true.

As it is known from modal logic, if something is necessary, then it is possible.

$$\text{nec } A \to \text{pos } A$$

Hence it can be assumed that the two signal perceptions are closely related. Rips and Heit and Rotello showed that this perception cannot be placed on a single scale. In the present research, we do not examine the difference between induction and deduction, but we borrow the experimental paradigm and apply it to a more subtle psycho-logical phenomenon. We know that in propositional logic, there are two different logical frameworks. The first one is the classical, which is based on the well-known truth-table method. The second is the so called intuitionistic logic, which can be considered as a kind of algorithmic logic. We are not only interested in whether the conclusion is true when the premises are true, but also whether the truth of the conclusion *can be recognized as true* by some algorithm or explicit method. Returning to the example (CMT), the argument is sound, because in this case Paul *cannot be not A*. If Paul were not A, he would not be B, but we know that Paul is B, so it is impossible for Paul to not be A. We cannot come up with any evidence to suggest that Paul is not A, because in that case we would find ourselves in a contradiction. However, we have no evidence to suggest that Paul is A. Intuitionistic logic follows the so-called Brouwer–Heyting–Kolmogorov interpretation of logic, according to this, propositions represent problems and when we assert a proposition, we assert that we know a solution to the problem represented by the proposition. So, even if we have no proof that A is false, it does not mean that we have a proof that A is true. Since this phenomenon is very similar to that of modal logic, the same paradigm can be applied.

$$\text{intu } A \to \text{class } A$$

2.1 Signal Detection Theory

Signal Detection Theory (SDT) is a widely used method in psychology. It is mainly used in two-alternative forced choice experiments, like the subject of this thesis, where participants are presented with signal and noise trials and during each trial or question they have to decide whether they are facing noise or signal. Essentially, there are four possible outcomes to these trials, hit, miss, false alarm and correct rejection. A hit occurs when a participant is presented by a signal and they correctly detect it. A miss is when they incorrectly identify a signal as noise. A false alarm is when a participant mistakes a noise trial for signal. And finally, a correct rejection is when they correctly identify a noise trial as noise. These outcomes can be conceived as a 2X2 table. It is important to note that it is enough to only account for hits, false alarms and the number of noise and signal trials presented to the participants in order to completely describe the data. And since it is the common procedure as well, this thesis only accounts for these pieces of data too. Both types of trials can be represented as values on a uni-dimensional strength scale, as both signal and noise trials produce strength values derived from Gaussian distributions along this same dimension. The signal is assumed to be stronger than the noise on average, so the mean of the strength value of the signal is greater than that of the noise, while their variance is the same. The participants are presented with yes or no questions during these trials and the assumption of the SDT is that the participants compare the strength of the currently presented trial with a fixed criterion. If the trial's strength is greater than that of the criterion then the participant deems the answer to that given question to be "yes", otherwise they would choose "no".

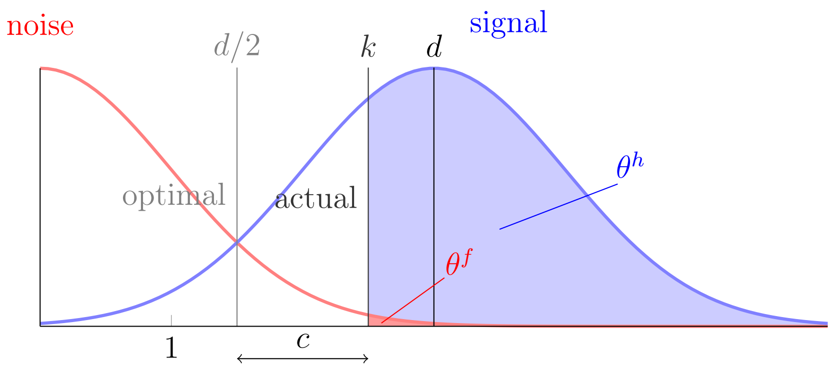


Figure 2.1: Equal variance SDT. The noise trial has a Gaussian$(0, 1)$ and the signal trial has a Gaussian$(d, 1)$ distribution.

Figure 2.1 represents the basic model of SDT. Since the strength scale has arbitrary units, Lee and Wagenmakers' advice is to set the mean of the noise to zero. [14] The mean of the signal is labelled d, which is a measure of the discriminability of the signal trials from the noise trials, since it is the distance of the means of the two distributions. *d/2* is especially important because it indicates the strength of an unbiased criterion, which means that a participant with a fixed criterion of *d/2* strength is the most likely to correctly classify trials when the frequency of noise and signal trials are equally likely. With k referring to a biased criterion, the distance of k and $d/2$ indicates the extent of the examined participant's bias and this distance is denoted c. If c is positive, then the participant is biased toward answering 'no' and if it is negative, then they are biased toward answering 'yes'. The former resulting in an increase in correct rejections at the cost of an increase in misses, while the latter causing an increase in hits at the price of an increase in false alarms. The area denoted θ^h is the hit rate and θ^f is the rate of false alarms, given the criterion of a given participant is k.

2.2 Statistics

In experimental science, holding as many factors constant as possible would be preferable, however, in psychology it is just not a realistic expectation. Lu and Rouder [15] argue that linear models, like the t-test, react to undesired variability in quite a negative way. Unintended variability from a simultaneous selection of participants and items tends to result in overestimation of confidence intervals and the inflation of Type I error in conventional analysis. And, unfortunately, this inflation is surprisingly large, and there can occur a potential lack of power of the by-subject and by-item analyses in psychological experiments, as demonstrated by Baayen, Tweedie and Schreuder. [16] Lu and Rouder explore a number of potential solutions to mend this issue, they advocate item-analysis, both item and participant analyses, a quasi-F statistic that accounts for item variability and a mixed linear model that accounts for both participant and item variation. Some of these techniques yielded positive results, however, they conclude that the more intuitive approach would be repetition. There is a significant drop from one experiment resulting in Type I error to the probability of two independent experiments resulting in Type I error. Yet, there is one more significant drawback of linear analyses, they are simply not as realistic as their non-linear counterparts. The use of non-linear statistical analyses have proven to be very popular among psychologists, and rightly so, but they don't come without their own problems. The most notable one being unmodeled variability, which can lead to distorted parameter estimates that can result in asymptotic bias, which is a major cause for concern, and one that an increase in sample size cannot attend to. Aggregating data seems to be the underlying cause for this distortion, therefore Lu and Rouder propose the solution to model both participant and item variability simultaneously, the former of which was something that Heit and Rotello chose to neglect in their research. Lee [13] further details this shortcoming of Heit and Rotello's research in the sense that it can be a major point of concern that they did not account for any individual differences because it is highly improbable that the participants had no variance in their capabilities regarding a higher order cognitive process, such as reasoning. Lu and Rounder along with Lee chose to tackle this issue by estimating a distribution of individual participant discriminabilities and biases, using Bayesian Hierarchical models, and in this thesis, I am going to apply that same technique as well.

2.3 Bayesian inference

Bayesian inference is a widely used statistical technique in psychology and cognitive science, it offers a principled and extensive approach for matching psychological models to data. Lee presents a number of significant and influential psychological experiments, such as the previously mentioned analysis of Signal Detection Theory of decision making, along with multidimensional scaling models of stimulus representation and the generalized context model of category learning to which he applies an array of Bayesian statistical methods. Through these demonstrations he illustrates the potential of Bayesian inference's capabilities to answer some important theoretical and empirical questions in an easy, coherent and comprehensible manner.
The basic principle of Bayesian inference is a combination of the uncertainty of a certain belief, that is the prior distribution, and the updating of that prior information through the use of observed data, thus producing a new updated piece of information, the posterior distribution. During a questionnaire, we cannot directly observe the ability of a participant θ, hence we need to treat this ability with a certain degree of uncertainty, which is the above mentioned prior distribution $p(\theta)$, this distribution can range from a completely uninformative uniform distribution to some very specific informative distributions. After establishing a prior distribution, we observe some kind of data D, in the case of this thesis, it is the results of the

questionnaire. And we update the prior θ in accordance with the newly introduced data, resulting in a posterior distribution *p(θ|D).* This distribution expresses the uncertainty of the participant's ability θ, assuming that the observed data D is true. Essentially, the posterior distribution is a combination of what we knew before seeing the data and what we've learnt from said data, thus the process of Bayesian updating prompts a reduction in the uncertainty of the value θ.

Bayes' theorem (2.1) provides a mathematical method to combine the information gathered from the data, *p(D|θ)*, which denotes the likelihood of D data occurring, provided that the ability of the observed participant is θ, with the prior distribution *p(θ)*, yielding a posterior distribution, *p(θ|D).*

$$p(\theta|D) = \frac{p(D|\theta)p(\theta)}{p(D)} \tag{2.1}$$

In 2.1 *p(D)* denotes the probability of data D occurring. It is the marginal likelihood of the applied model, also known as the evidence, which is a single number that is independent of the participant's ability, θ; *p(D)* ensures that the area below the posterior distribution adds up to 1, otherwise it would not qualify as a distribution. According to Lee, *"...the marginal likelihood measures the average quality of the predictions that a model has made for the observed data. The better the predictions, the greater the evidence."* [14]

The marginal likelihood can be computed by following the concept of the Law of Total Probability, which requires calculating the likelihood for each parameter value, weighting them by their prior probabilities and adding them up. In the case of a continuously varying parameter θ, such as the one this thesis focuses on, the summation is replaced by an integral, shown in (2.2).

$$p(D) = \int p(D|\theta)p(\theta)d\theta \tag{2.2}$$

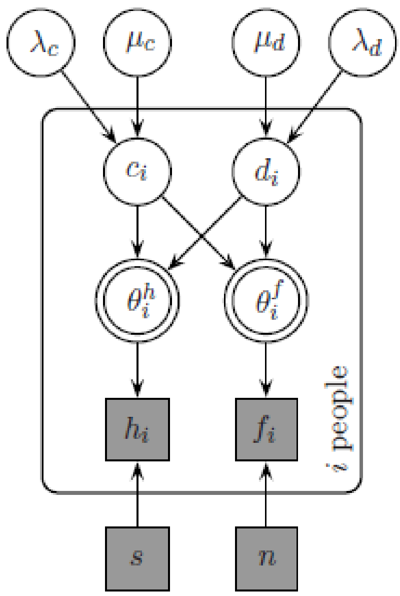
2.4 Markov chain Monte Carlo method

According to Lu and Rouder, one of the main drawbacks of non-linear hierarchical models is tractability; they proved to be exceptionally difficult to implement. Lee argues that for the longest time, only relatively simple models were possible to assess because of how complicated the evaluation of the posterior distribution can be. Gilks, Richardson and Spiegelhalter [17] argue that the complexity comes from the integration in the normalization constant, the marginal likelihood, which is apparent from (2.2). But this landscape was completely transformed by the emergence of computer-driven sampling, known as Markov chain Monte Carlo(MCMC). The popularity of Bayesian inference in psychological sciences skyrocketed because the main obstacle of this method has been lifted. And in the statistical analysis of this thesis, I apply an MCMC method as well, following the suggestions of Lu and Rouder and Lee.

Following Lu and Rouder's suggestions, I used a hierarchical version of Signal Detection Theory in this thesis, outlined by Lee, in order to account for individual differences among participants regarding discriminability and bias.

2.5 A Hierarchical Extension of SDT

Lee and Wagenmakers suggest the following hierarchical model to analyze the data gathered by Heit and Rotello:



$$
\begin{aligned}
\mu_d, \mu_c &\sim \text{Gaussian}(0, .001) \\
\lambda_d, \lambda_c &\sim \text{Gamma}(.001, .001) \\
d_i &\sim \text{Gaussian}(\mu_d, \lambda_d) \\
c_i &\sim \text{Gaussian}(\mu_c, \lambda_c) \\
\theta_i^h &\leftarrow \Phi(\tfrac{1}{2}d_i - c_i) \\
\theta_i^f &\leftarrow \Phi(-\tfrac{1}{2}d_i - c_i) \\
h_i &\sim \text{Binomial}(\theta_i^h, s) \\
f_i &\sim \text{Binomial}(\theta_i^f, n)
\end{aligned}
$$

Figure 2.2: Graphical model for hierarchical signal detection theory (Lee and Wagenmakers 2014)

This model, illustrated by 2.2, uses SDT to infer discriminability (d_i) and bias (c_i) for the ith participant, who has the hit rate of θ_i^h and false alarm rate of θ_i^f. We know how many hits and false alarms each participant scored; for the ith participant these are h_i and f_i respectively, and these scores have a binomial distribution with the probability of the two above mentioned rates and an s number of signal trials in h_i and an n number of noise trials in f_i. Individual differences are accounted for through this hierarchical model with the assumption that the individual biases and discriminabilities come from Gaussian group-level distributions with the means of μ_c, μ_d and the precisions of λ_c, λ_d respectively, which is a fair assessment of Lee's, considering that the overwhelming majority of intelligence related statistics show a normal distribution, and since both the bias and the discriminability are high level cognitive attributions, they are more than likely to show such a distribution as well. [18]
θ_i^h and θ_i^f are also indicated in 2.1, these are the highlighted areas below the bell curves of the signal and the noise, with the cumulative distribution functions for θ_i^h being:

$$\Phi\left(\frac{1}{2}d_i - c_i\right)$$

And for θ_i^f it being:

$$\Phi\left(-\frac{1}{2}d_i - c_i\right).$$

Lee's results can thus be represented by a two-dimensional classification in which the signal recognized by the human brain (in this case, whether the inference is valid or not) is evaluated according to two parameters. The first parameter describes how easy it is to separate the signal from the noise by the subject (d), and the second one describes how courageous they are to do so (c). The figure shows the approximate posterior distribution of the latent variable pair (μ_d, μ_c) obtained by Bayesian interference. These are the inferred means of the normal parameters d and c, respectively.

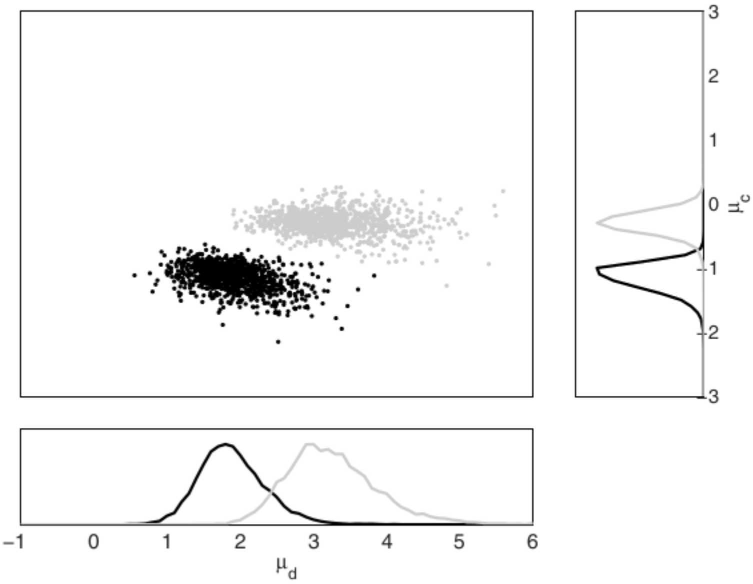


Figure 2.3: Inductive reasoning is marked as black, deductive as gray. Deductive arguments are more effectively ($\mu_d^{\text{ded}} \approx 3/\mu_d^{\text{ind}} \approx 2$) and inductive ones more discouragedly ($\mu_c^{\text{ded}} \approx -0.5/\mu_c^{\text{ind}} \approx -1$) recognized by the subjects of the experiment. (Lee and Wagenmakers 2014)

The result can also be well represented by a diagram depicting the normal curve of the signal and noise in the SDT paradigm, so that the curves corresponding to the two signals are displayed according to the average values of the discriminability of deductive and inductive reasoning. The figure shows that, according to Heit and Rotello's data, the inferred mean discriminabilities assume two well-separated signals. The first of these (the induction one) differs by only 2 standard deviations from the noise, while the one for deduction differs by 3.5. The type I error is almost around the conventionally considered magnitude of significance.

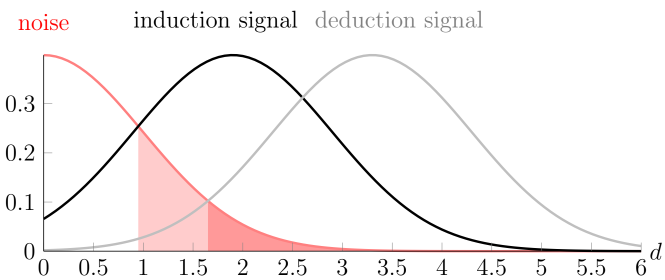


Figure 2.4: The magnitude of the type I error is way smaller in the recognition of deduction than in the case of induction.

Chapter 3

Methodology

The two main aims of this thesis are to replicate the experiment conducted by Heit and Rotello, with the statistical alterations suggested by Lee, and find further evidence to support either one or two-process accounts of reasoning, and the other aim is to see whether teaching logic to high school students using the Coq proof assistant environment is more efficient and leads to a deeper understanding than teaching it using the coursebook and strictly following the recommendations of the Hungarian National curriculum. In this research, we focus only on deductive reasoning. What in the cognitive experimental paradigm will be deduction and induction will now be classical logical deduction and intuitionist logical deduction. The initial plan was to teach two groups, one with Coq and one using conventional methods. I allocated four-four classes to teach both groups their corresponding curriculum, and they filled in a questionnaire at the very first logic-class and they filled in the same questionnaire at the end of our last class. Both of these groups were from a bilingual school, where I have been teaching part-time and I had been teaching both groups for a month and a half at the time of the research, so I had some knowledge of their skills and personalities, which made this research go much smoother. They have two Hungarian and two Spanish mathematics classes a week, which causes some difficulties, since the two mathematics classes focus on different topics and it is often very difficult to build on knowledge that the either learnt or was supposed to learn at Spanish mathematics classes, since they often do not know the names of mathematical concepts and theorems in Hungarian and vice versa, and on top of that Spanish mathematics teachers also struggle a lot with the initial language barrier. I have been teaching them Hungarian mathematics. The conventional methodology group was one half of a 10th grader class, and the Coq group was a half of a 9th grader class; there were 20 and 19 students in these classes respectively. Both of these groups are one year older than usual, since they all had a 0th year as a language preparation year, so the 9th graders were 15-16 years old and the 10th graders were 16-17 years old. There was one student in the conventional group, who didn't speak Hungarian and since this topic is so heavily language oriented, I didn't include her in my research, and she got some individual tasks to work on during these classes. Both of these groups proved to be very cooperative and patient throughout my research, and most of them take school quite seriously and they are very grade-oriented too. Since this school is considered relatively good, the parents tend to expect a lot from their children, which pushes them very hard, often too hard even These were all very able students with exceptionally high language aptitude, the latter of which translated surprisingly well to learning logic; although mathematics was not their forte, they were trying their best and performed remarkably well. All in all, the ability, the attitude and the behaviour of the students did not pose any obstacles, quite the opposite, they made many aspects of the research quite easy and free-flowing.

3.1 Questionnaire

The questionnaire contained twelve sets of statements with a conclusion each. The students' job was to decide whether those conclusions were valid or not and to rate on a scale from one to five how confident they were in their answers. These sets of statements ranged from basic modus ponens and transitive law to the much more complex Peirce's law and Contraposition. Since the Coq group was learning intuitionistic logic and the conventional group was learning classical logic, it was only natural to evaluate their answers from both perspectives, focusing on the results of the discussed system of logic for each group. The statements were all stripped of their meanings, so the students could not rely on any background knowledge. This would have been detrimental in more than one aspect, as Heit and Rotello discuss the statement *"Jill rolls in the mud, therefore Jill rolls in the mud and Jill gets dirty"* [11], it has a potential for participants to make the immediate association between mud and dirt, thus finding this conclusion valid, whereas it is not, since there is no logical connection provided between mud and dirt in the premises and in such a case the participants would only rely on their background knowledge. And in the case of the following statement *"Jill rolls in the mud and Jill gets clean, therefore Jill rolls in the mud"* [11] has the opposite effect, it is deductively valid but the participants' background knowledge can override their decision-making, thus leading them to deem it invalid. The first example is deductively invalid but inductively strong, whereas the second one is deductively valid but the students could evaluate it inductively weak, which contradicts one of the testable predictions of the criterion-shift account. This prediction states that the relative ordering of two arguments should be the same regardless of whether they are judged based on their deductive correctness or inductive strength, thus deductive validity would presuppose that the statement is inductively strong as well. Heit and Rotello argue that this contradiction merely stems from the introduction of additional contextual information, which activates the participants' background knowledge, which disturbs the process of reasoning. And because the process of reasoning is disturbed by an external force, Heit and Rotello argue that such an example is not a conclusive piece of evidence against the account. So they advise researchers to omit any possible triggers of background knowledge from a questionnaire, which I abided by in mine.

3.2 Teaching logic using Coq

With the Coq group, we spent all of these four classes in the computer room and we started our first class with a brief summary of what a logical statement and a conclusion are. We discussed two example sentences, one was a very basic modus ponens sentence, to demonstrate how a valid conclusion works. The other one was a statement with an invalid conclusion, which was very similar to one of the previous examples: "Jill rolls in the mud, therefore Jill rolls in the mud and Jill gets dirty". Then after we had discussed what statements are and what a valid and an invalid conclusion looks like, students got 15 minutes to fill in the questionnaire.

This first questionnaire, just like the conventional group's first questionnaire, was conducted through a quiz website that students could access through their phones, called Quizizz.com. I had used this website numerous times before during my English as a second language classes, so I was familiar with the layout and all the features. The reason I decided to do it this way, instead of doing it on paper was, so I could measure the students' response time as well as their answers, which would have provided some valuable information to support or question the theory of there being two kinds of reasoning. During the four classes of the research I decided to discard the data on the students' response time because I deemed it unreliable, since this website can take some time to display a question, and also some phones might be slower than others, thus resulting in a distortion of the response times measured. The end-of-course questionnaires were filled in on paper in both groups. Measuring response time could have produced valuable data, and it could be a great follow-up for this research but it would need to be carried out in a controlled and reliable manner, unlike the one I tried.

After the quiz, we started to get familiar with the Coq environment, we used an online version of Coq (https://jscoq.github.io/scratchpad.html) which worked flawlessly, in fact, it proved to be even more suitable for my teaching purposes than the desktop version. Since every proof that I wanted to showcase had quantifiers and logical operators in them, it was very beneficial for us that this online version automatically replaced them with their corresponding mathematical symbols, so for example, the syntactic element of "forall" was replaced by the $\forall$ symbol. It made it easier for the students to remember the underlying concepts of these quantifiers and operators because these visual aids provided further reinforcement to their learning, just as a realia or a flashcard would in a language learning environment.

At first, it was quite challenging for them to simultaneously learn the syntax of Coq and also grasp the thought process of proving and the structure of a logical statement. But since both are at a very similar level of abstraction, these two seemingly separate learning processes ended up being heavily intertwined, and thus one helped reinforce the other. Therefore this initial challenge turned out to be quite an advantage in the long run and judging by the results of the questionnaires, it proved to be more than worthwhile and would have yielded even more advantages, had the course lasted longer. Another helpful aspect of the usage of Coq's syntax was that these students have exceptionally high language-learning aptitude, so despite the

early difficulties, they tackled the language learning side of these classes with ease, which, on the one hand, gave them a sense of accomplishment straightaway, and on the other, it helped them get a better grip on the abstraction of logic itself.
Generally speaking, providing students with a familiar environment in learning is always beneficial, and in the case of logic, the introduction of linguistic elements comes naturally and fits seamlessly into the concepts of operators, quantifiers and mathematical abstraction, thus helping students with high language-learning aptitude, such as the ones from bilingual schools, understand these concepts much more easily. And Coq proved to be a clean and accessible environment to aid this part of the learning process.

3.3 The Coq course

We started the course by proving the transitivity of a set of implications.

If Panni is A, then Panni is B.
If Panni is B, then Panni is C.
Panni is A.

Panni is C.

First, we turned the logical statement into Coq code, during which, we showcased a number of logical and linguistic elements as well as the formalization of Coq.

```
Theorem Chain_Rule:
forall A B C : Prop, ( A /\ ( A -> B ) /\ (B -> C ) ) -> C.
```

Here we had a great opportunity to get familiar with the logical quantifier of $\forall$ and the logical operator $\wedge$, both of which were novel concepts for these students. After entering Proof mode, we used the `'intros'` command, which introduces the universally quantified variables, in our case `'A'`, `'B'` and `'C'` and the premises, which in this case was the proposition $(A \wedge (A \Rightarrow B) \wedge (B \Rightarrow C)) \Rightarrow C$.
This proposition unsurprisingly proved to be quite complex for the students at first, since it contained many new operators in this new environment. In order to make it much more manageable, we used the `'intros F'` command once more to break down the proposition into the premise $(A \wedge (A \implies B) \wedge (B \implies C))$, which was

labeled 'F' and the conclusion C. Then we further disassembled the proposition F, using a new command 'destruct' twice in a row:

```
destruct F as [F1 F2].
destruct F2 as [F2 F3].
```

After all this, we were presented with the following in the goal window:

```
1 subgoal
A, B, C : Prop
F1 : A
F2 : A -> B
F3 : B -> C
______________________________________(1/1)
C
```

And this line of statements is in the exact same form that we previously encountered during the questionnaire. Now that we had established how to formulate logical statements in the Coq environment, it was time to discuss how to prove them using intuitionistic logic. Intuitionistic logic is a branch of logic, where a statement is valid, if it can be proven but it does not include the law of excluded middle and double negation elimination, so we could not rely on truth tables; only on the process of constructive proofs, which is exactly the same as Coq's inner logic. In the goal window, Coq always tells the user what kind of statement the process of deduction requires next, in this instance, it was the conclusion of the whole chain itself: 'C', so the students' job was to find an implication that has 'C' as the the consequent, and in this example 'F3' fits that role. At first, students needed quite a bit of time and effort to grasp the notion of this method of proving, especially considering that they had not had much experience with mathematical proofs at all, let alone one that is in such an abstract form and features so many novel linguistic features. One thing that really helped them during this proof was to find suitable substitutions for the variables: 'A', 'B' and 'C', in our case, I used the following substitutions: 'A': bug, 'B': insect, 'C': arthropod; evoking the well known phrase: "Every bug is an insect", with the added layer of "Every insect is an arthropod.

After some deliberation and many explanations, the students slowly got the hang of the process of proving, so we started to apply the given propositions one-by-one. In Coq, it was as straightforward as using the `'apply'` command along with the appropriate statements. First, we used the `'apply F3'` command, then the Coq goal window was asking for the variable `'B'`, so we used the `'apply F2'` command, which had `'B'` as the consequent, then finally Coq was asking for the variable `'A'`, so we used the `'apply F1'` command; this last one could have been replaced by the `'exact F1'` command as well. One of the main advantages of using Coq for this proof is that it showcases the incremental nature of this proof, that we had to apply the statements one after the other, like a chain. The reason I chose the transitivity rule as the introductory proof is because of its simplicity on the one hand, and because, for students, it seems very daunting at first how long winded and complex the formal phrasing of the statement is, but as it gets broken down into simple singular inductions, it gets apparent how simple the core concept actually is and how straightforward the proof itself is. During the last 10 minutes of the class, students, individually, started working on the proof of the long chain rule, which is the following code in Coq:

```
 Theorem longchain: forall A B C D : Prop,
( A /\ ( A -> B ) /\ ( B -> C ) /\ ( C -> D ) ) -> D.
```

It has the same proof as the previous one with the only added steps of an additional variable and another `'destruct'` and `'apply'` command. Some students, unsurprisingly, struggled with the application of the steps of the proof at first and we even ran out of time, but after some work everyone managed to finish the code by roughly the 10th minute of the second class, which seemed a bit discouraging to me at first, but as we progressed through the course, I slowly realized that this rough start actually helped us in the long haul. The next proof on our agenda was the Modus Tollens, where we introduced a new logical operator, the negation: ¬, along with its implementation in Coq. The theorem in Coq is the following:

```
Theorem modus_tollens : forall P Q : Prop, ((P -> Q) /\ ¬ Q) -> ¬ P.
```

The proof starts the exact same way as before, with the `'intros P Q'` and the `'intros H'` commands, followed by the `'destruct H as [H1 H2]'` command.

These introduce the variables, the proposition on the left side of the implication and break down the conjunction in the newly introduced proposition respectively. After all this, comes a new command, the `'unfold not'` command, which replaces a defined term with its definition, in this case it replaces the `¬ P` with the implication $P \rightarrow False$, which is the intuitionistic definition of 'not', with the only caveat being that instead of $False$, we would use the term absurdity ($\perp$), which is important to note because unlike in classical logic, we can't just assume that $\neg P$ is just the complement of P, since the law of excluded middle is not included in this logic. After applying the `'unfold not'` command, we used the `'intros H3'` command, and the justification for this step was that the previous command resulted in the introduction of an implication in the consequent, thus presenting a new antecedent that was possible to add to the list of propositions on the left side of the main conclusion. Then we repeated the same `'unfold not in H2'` command, with the only difference that it affected the `'H2'` proposition instead of the consequent of the entire statement. And finally, after all this, we were presented with the following in the goal window:

```
1 subgoal
P, Q : Prop
H1 : P -> Q
H2 : Q -> False
H3 : P
______________________________________(1/1)
False
```

This is strikingly similar to the proof of the previously discussed chain rule and after some deliberation the students realized it as well, and were happy to break down the chain step-by-step, by linking the propositions together, like this: `'apply H2'`, then `'apply H1'`, and finally `'apply H3'`.

During the last bit of this class and the first half of the third class, students had the opportunity to practice what we had learned on the following proof:

$(P \wedge Q) \rightarrow \neg(P \rightarrow \neg Q)$

Unlike before, I let the students try and translate this theorem into the language of Coq by themselves, with me, obviously, going around and helping them individually.

It led to some difficulties, but all in all, the overwhelming majority of the class was starting to get a grip of the notion of Coq and most issues only stemmed from typos and minor syntactic errors, which were easy to mend. The majority of the class managed to get most of the proof done by themselves, with only little help from me. They introduced the variables and the left side of the proposition, replaced the $\wedge$ and $\neg$ operators with their definitions and broke down the implications as well. They only hit a roadblock after they encountered the following text in the goal window after breaking down a three part implication:

```
2 subgoals
P, Q : Prop
F0 : P
F1 : Q
______________________________________(1/2)
P
______________________________________(2/2)
Q
```

Here Coq was expecting us to solve two subgoals instead of only one, which was, expectedly, a daunting task for the students at first, but thankfully, proving it was just as simple as linking a chain of propositions one after the other, just like before. They only needed to use the `'apply F0'` and then the `'apply F1'` commands.
The majority of the third class was centered around the logical operator 'or' ($\vee$), how to break it down during a proof and the proof tactics of finding contradictions. After a brief theoretical overview of the operator $\vee$, we observed how it behaves in a proof in the Coq environment. We proved the following logical theorem:
$(A \vee B) \rightarrow (\neg A \rightarrow B)$
The proof starts as usual with the introduction of the variables and the propositions wrapped up in the antecedent. The usual first step after this had been to apply the `'destruct'` command to as many propositions as possible and, we did exactly that here too. The novel aspect of breaking down a proposition containing the operator $\vee$ is that it splits the proof into two parts, one where we have A among the propositions and one where we have B there, and just like before, it is expressed in Coq as having two subgoals in the goal window. Although it is not necessary, it proved to be a good

idea to use the '-' tactic to unfocus one of the subgoals, which does not alter the proof at all but makes it a bit easier to follow. After unfocusing one of the subgoals, we were presented by the following goal window:

```
1 subgoal
A, B : Prop
H : A
H0 : ~ A
______________________________________(1/1)
B
```

And here is where we touched upon the other new concept, which was the 'contradiction'. The students seemed to have a much easier time comprehending this concept, which might have just been due to them having more experience with proofs, but it was reassuring nonetheless. After using the `'contradiction'` command, this subproof was complete, and all we had to do was focus on the other subgoal, using the '-' tactic once more, where `'B'` was among the propositions, and there was just a simple `'apply'` command needed to finish the proof.

The second half of the third class and the first half of the fourth class were allocated to individual practice, here students proved the following:
$(A \wedge B) \rightarrow \neg(\neg A \vee \neg B)$
This proof did not require anything new, all the students had to do was apply the same techniques as before. And even though they encountered some difficulties, with some help, they could eventually tackle the problem. This proof is very similar to the previous one: it starts with the usual `'intros H'` command, followed by the `'destruct H as [H1 H2]'` command, which replaces the conjunction with its definition, hence producing `A` and `B` as two separate propositions. The biggest hurdle came after applying the `'unfold not'` command, which produced the following in the goal window:

```
1 subgoal
A, B : Prop
H1 : A
H2 : B
```

```
______________________________________(1/1)
(A -> False) \/ (B -> False) -> False
```

The reason so many students got stuck here was because they got overwhelmed by the large number of operators appearing. Fortunately, the ones who were not afraid to experiment, quickly realized that by applying the tried-and-tested sequence of the `‘intros H’` command, followed by the `‘destruct H’` command, it got much simpler:

```
2 subgoals
A, B : Prop
H1 : A
H2 : B
H : A -> False
______________________________________(1/2)
False
______________________________________(2/2)
False
```

Here, all they had to do was apply a chain of propositions one after the other, like so: `‘apply H’` then `‘apply H1’` to prove the first subgoal, then `‘apply H’` and `‘apply H2’` to complete the second subproof. Just like before, the addition of the ‘-’ tactic could make it a bit easier to comprehend, since it saves students from information overload.

The second part of the fourth class focused on the differences between the universal (∀) and the existential (∃) quantifiers and their relations to one another. Here I gave them four statements and their job was to find out which implies which. These four statements were the following: “Everybody loves somebody”, “Everybody is loved by somebody”, “There is somebody who loves everybody” and “There is somebody who is loved by everybody”. It took them some time to understand these relations, but they enjoyed this change of pace very much, and even the ones who struggled with the syntax of Coq and the underlying logic of proofs, had a lot of success with this task, and it proved to be a good final note to this course. After this conclusion, all that was left was to fill in the questionnaire once more.

Overall, this Coq course proved to be a success, most of the students seemed to really enjoy it and changing the learning environment motivated them quite a bit too. The fact that they had so much freedom in how they wanted to tackle a proof and that they were encouraged to experiment as much as they wanted to, had them much more engaged than usual. It is apparent that there were a lot of difficulties too and allocating just four classes to this topic was far too short to delve deeper into Coq and proofs in general. But even this very surface level course managed to yield some promising results, and most notably, managed to engage the students' attention, which, in itself, gives ample reason to be optimistic and open-minded about the incorporation of computers and Coq in teaching logic.

3.4 Teaching logic the conventional way

The second observed group focused on learning classical logic in a very conventional and old fashioned environment, without implementing any digital learning materials. With this group, my main aim was to be as faithful to the aims and recommendations of the Hungarian National curriculum. It states that the students have to be able to decide whether a statement is true or false, they have to be able to use negation regarding basic logical statements, they have to understand the logical concepts of implication ($\rightarrow$), equivalence ($\iff$), the operators 'and' ($\wedge$), 'or' ($\vee$) and 'exclusive or' ($\underline{\vee}$), and they need to be able to use them as well, and finally they need to be able to use the universal and the existential quantifiers properly. The national curriculum puts a lot of emphasis on the parallels between classical logic and set theory, and focuses on the connection between the operators in logic and in set theory. This thesis in no way aims to comment on either the Hungarian National curriculum or teaching classical or intuitionistic logic, these are only used as instruments to compare and contrast two teaching methods in the light of some influential psychological research. One additional equipment of this research is the coursebook, we used the 'Matematika 10. - Az érthetõ matematika' (ISBN 978-615-6256-35-5) as our coursebook, which is a certified mathematics coursebook for high school students, the content and structure of which are in accordance with the Hungarian National curriculum of 2020.

Some of the main features of classical logic is the following: The law of excluded middle states that for every proposition, either the proposition itself or its negation is true.The principle of bivalence, which states that every proposition is either true or false but cannot be both. Truth functionality, which states that the truth value of any complex statement built up following the syntactic rules of the system is unambiguously determined by the building blocks of said complex statement. Double negation states that $(\neg\neg A)$ and A are truth-funcionally equivalent, which means that their truth tables are identical. The most useful tool that classical logic provides students with is the truth table, which gives them a way to evaluate the truth value of any given logical statement. [19]
There are many arguments for and against the universality of classical logic. Many suggest that it is the ideal environment for guiding reasoning and thus it is "the one right logic"; it is very simple yet exceptionally effective and strong. But there are also many who oppose this distinguished role of classical logic, arguing that there is no reason, aside from that of historical significance, to put classical logic on a pedestal ahead of other logics, for example intuitionistic logic or relevance logic. [20]. This thesis does not aim to weigh in on that question but it is important to note that the position of classical logic in mathematics is far from uncontested, and including it in the Hungarian National curriculum is not a necessity but a choice made by the authors of the Hungarian National Curriculum.

3.5 The Classical Logic course

Just like the Coq group, this group had four classes allocated to learning logic but in this case, the focus was on classical logic, using conventional methods. Unlike the Coq group, this group consisted of 10th graders, so they may have had some advantage in terms of their cognitive maturity, although the statistics show very little difference in their performance regarding the first questionnaire, but most importantly, similar to the Coq group, they had no prior explicitly logic-focused studies. The first class started identically to the Coq group, we discussed what a valid and an invalid conclusion looks like through a couple of examples, one was a modus ponens sentence to demonstrate how a valid conclusion works and the other was an invalid conclusion that was taken from Heit and Rotello, which was the previously mentioned "Jill rolls in the mud, therefore Jill rolls in the mud and Jill gets dirty". After this short demonstration of logical reasoning, students had 15 minutes to fill in the questionnaire, just as the Coq group did.

After the questionnaire we discussed the foundations of classical logic, focusing on the law of excluded middle and the principle of bivalence. Then we started discussing the 'and', 'or' and the 'exclusive or' operators and how they differ from their everyday use. We practiced these through a few given statements found in our coursebook, all of which touch upon some mathematical statements, which, in this book, not only aims to teach logic but they also act as a means to revise some of the theorems and definitions learnt the previous year. It is not necessarily a negative thing to use logic as a device for revision but it indicates that it is not in the forefront of Hungarian high school mathematics. The syllabus of this coursebook only allocates two classes for logic, so in many cases, I had to provide additional material for the students, and I did so in this case as well, I gave them a number of real life examples to demonstrate the concepts of these logical operators because I wanted to paint a better picture of logic, that it is not just simply a mathematical device but a mathematical field of study as well, which can stand on its own and does not only exist to just supplement other fields of mathematics. I did not deviate from the intended course of learning, recommended by the book, nor did I skip over any parts, I only expanded it.

The focus of the second class was the logical operators, their connection with set theory and their truth tables. The four operators that we discussed were the 'and', 'or', 'exclusive or' and the implication, in this order. The 'and' operator was relatively easy to understand for everyone, and in contrast to the Coq course, the 'or' operator proved to be quite straightforward as well, the students were very receptive to truth tables and the visual representations of these operators in the form of Venn diagrams. At first, it was a bit surprising to me how they were more open to learning logic the conventional way opposed to the computer aided one, but it turned out that it was not the learning environment that made the difference but the allocation of the course syllabus. The Coq course was much more difficult and alien towards the beginning because of the addition of Coq's syntax, whereas the conventional course's structure was more gradual and less demanding at the start, and it also provided students with a very easy-to-use technique to determine the truth value of any logical statement, the truth tables. However, it is important to note that the Coq course turned much more satisfying towards the tail end of the course and continuing it would have most likely been a better learning experience, whereas the conventional course turned a bit blander by the third class and there did not seem to be any desire or enthusiasm on the students' part to continue beyond the fourth class. Regardless of the allocation of the course material, the use of Venn diagrams very quickly and clearly conveyed the meanings of these concepts and the introduction of truth tables made it possible for students to compute, thus creating a familiar learning environment, where they felt comfortable. Although the lack of stimuli and the missing sense of exploration and experimentation led to a slow decline in their intrinsic motivation, which seemed like the biggest flaw of this course. These classes felt like run at the mill mathematics classes and they did not provide students with any new impulses or novel, memorable pieces of experience, despite my best efforts to make them as interactive as possible; however, the efficiency of these classes is undeniable and are suitable to convey as much information as possible in such a short time, and thus they clearly fit the demands dictated by the Hungarian national curriculum.

The third class was centered around the operator of negation; conceptually it proved to be quite easy to grasp because the visualization of Venn diagrams and example sentences made it much less abstract and truth tables provided a simple, yet effective method to handle them. The negation of the 'and', 'or' and 'exclusive or' operators did not cause much trouble but the negation of implication turned out to be quite counter-intuitive at first and no visualization or truth table helped with comprehending it. Thankfully, example sentences worked very well, my example sentence was the following excerpt from a well-known song: "Thunder only happens when it's raining" and the mathematical representation of it is as follows: 'Thunder' $\rightarrow$ 'Rain'; and the students' job was to come up with a situation that would prove that the singer of this song is wrong. The reason I chose this particular phrasing of an implication is because of how easy it is to negate it: "Thunder does not only happen when it's raining", thus there are situations when thunder happens but it is not raining, and from here on, it quickly started to click for most of the group during the practice tasks in the coursebook. The negation of implication proved to be one of the highlights of this course because, for many students, it was the first real eureka moment that came from them discovering something without me providing them any significant help.

The final class focused on the concepts of equivalence and contradiction, as well as the differences between the universal and the existential quantifiers. We defined equivalence through truth tables with the support of some example sentences, and then we briefly discussed contradictions and how we can use them in proofs. Finally, we touched upon the two above-mentioned quantifiers and we did the exact same exercise as we did with the Coq group, which was very well-received once again and proved to be a good finale to this course as well. Finally, we closed the course with filling in the questionnaire once more.

Chapter 4

Results

The directly measured results are shown in the diagram below. IL indicates the group in which I taught the Coq course, CL indicates the group in which I taught logic in the traditional way. The questionnaire was filled in at the beginning and the end of the course for both groups. In the case of the second questionnaire, I paid special attention, so that the students could practice inferences for at least 20 minutes in the first part of the lesson, before filling in the questionnaire. Let's see the results of the 9th grade high school group with the code IL first.

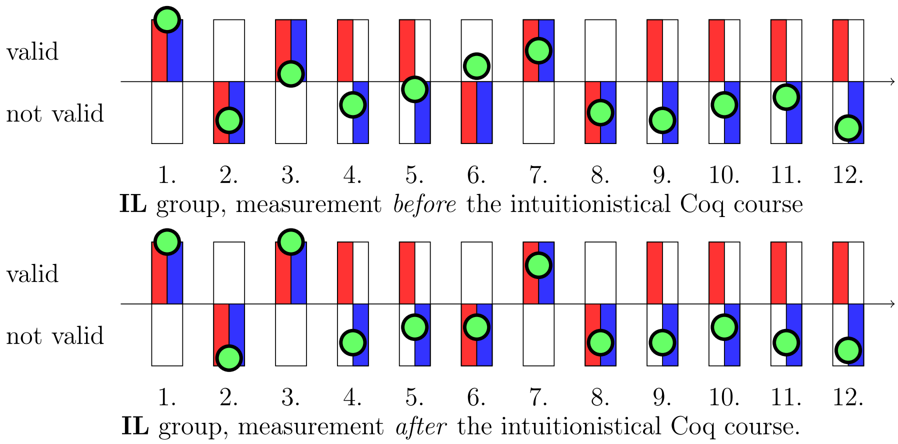
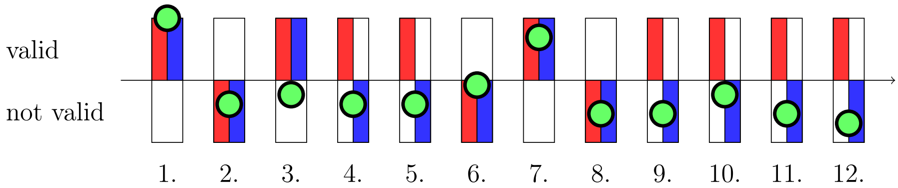
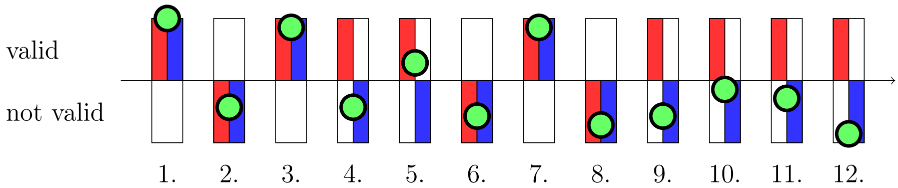


Figure 4.1: The IL group's first and second questionnaire. The ratio of the answers are the green dots. Classical evaluation is in red, intuitionistic evaluation is in blue.

As the diagram shows, the mode of the responses (except for one, in the first measurement) are all in the correct half, which means that the group in average provided correct answers in the intuitionistic way of logic (blue rectangles). The hit rate for the classical logic answers, partly because of how the good intuitionistic hit rate was, is quite low. Only 5 or 6 correct answers were given, out of the total of 12, where validity is drawn from the classical logical framework (red rectangles). These students have apparently never learned classical logic before. Of course, they did not learn classical logic after the first survey either, so they were not expected to know it at the time of the second survey. Interestingly, the result remains similar when we look at the group labelled CL.



The **CL** group, measurement *before* the classical traditional course.



The **CL** group, measurement *after* the classical traditional course.

Figure 4.2: The CL group's first and second questionnaire. The ratio of the answers are the green dots. Classical evaluation is in red, intuitionistic evaluation in blue.

The results are very similar, even though this group learned classical logic during the time between the two questionnaires.

Turning to the data calculated from the *generative graphical cognitive model*, I would first highlight some details regarding the background of it: the data was processed using the language R in the RStudio environment, with the help of the rjags, coda and LaplacesDemon packages. The hierarchical model I used is the one that was detailed by Lee and Wagenmakers, who present a Winbugs implementation of it, and their exact model was translated into JAGS, using the rjags package.

The figures below show the Bayes-inferred results of the joint probability distribution of the latent parameter pair $\langle \mu_d, \mu_c \rangle$, which are the means of the probability variables' objective signal discriminability d and bias c, resulted by the responses and the evaluation in relation to the two logics. The point clouds shown in the figure are the posterior of the latents generated in an MCMC method with a sample size of 10,000.

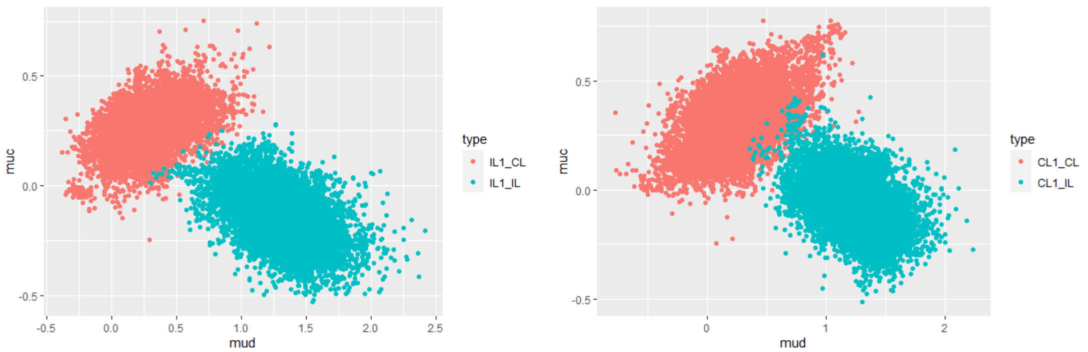


Figure 4.3: Left: the preliminary measurement of the IL group. Right: the preliminary measurement of the CL group. The red data cloud is the evaluation according to classical logic, the blue one is the evaluation according to intuitionistic logic. The two latent variable pairs $\langle \mu_d, \mu_c \rangle$ show very similar distributions.

It can be seen that the intuitionistic logic-evaluation has a larger mean of μ_d, i.e., assuming that the students chose between signal and noise. In relation to this logic, they were more able to distinguish signal from noise, however, it is clear that both μ_d values are very close to the standard deviation of the signal. This means that classical logic is barely recognizable to them and the non-classical is difficult to recognize too. Of course, I do not claim that students think in an intuitionistic way, but that it is statistically closer to their thinking than the classical one.

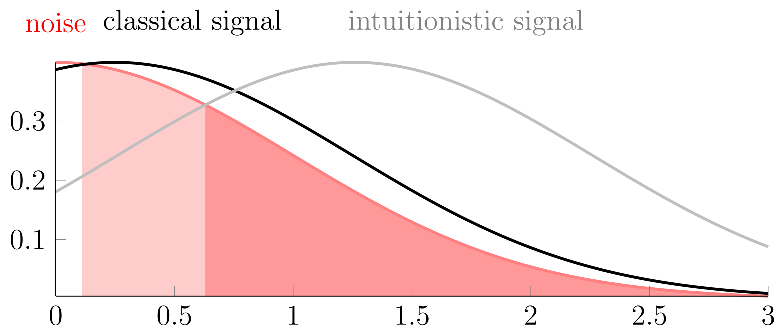


Figure 4.4: The means of the μ_d distributions and the signal locations of the two examined logics. The magnitude of the type I error was very large for both logics in the beginning of the courses.

In the absence of prior knowledge on the part of the students, it can be assumed that the two measurements provide the same distribution for the two groups. To verify this, a Bayesian model comparison was used. I examined the predictability of the CL data cloud by looking at the near-normal distribution of the IL data cloud as an alternative model and the null model was a non-informative prior, with a high standard deviation.

group	mean of μ_d classical	sd of μ_d classical	mean of μ_d intuitinistic	sd of μ_d intuitionistic
IL	0.2542	0.2587	1.2497	0.2354
CL	0.2446	0.2380	1.2677	0.24401

From these data, in the case of the first measurement, it emerged that the IL distribution strongly explains the CL distribution in comparison to the non-informative model.

signal	data model	null model	alternative model	BF_{10}	validity
int.	$\mathcal{N}(1.2677, 0.24401)$	$\mathcal{N}(3, 10)$	$\mathcal{N}(1.2497, 0.2354)$	29.9	strong
class.	$\mathcal{N}(0.2446, 0.2380)$	$\mathcal{N}(3, 10)$	$\mathcal{N}(0.2542, 0.2587)$	32.2	strong

Regarding the evaluation of the second measurement, we will see that descriptive analysis shows an improvement in discriminability, although not to the same extent.

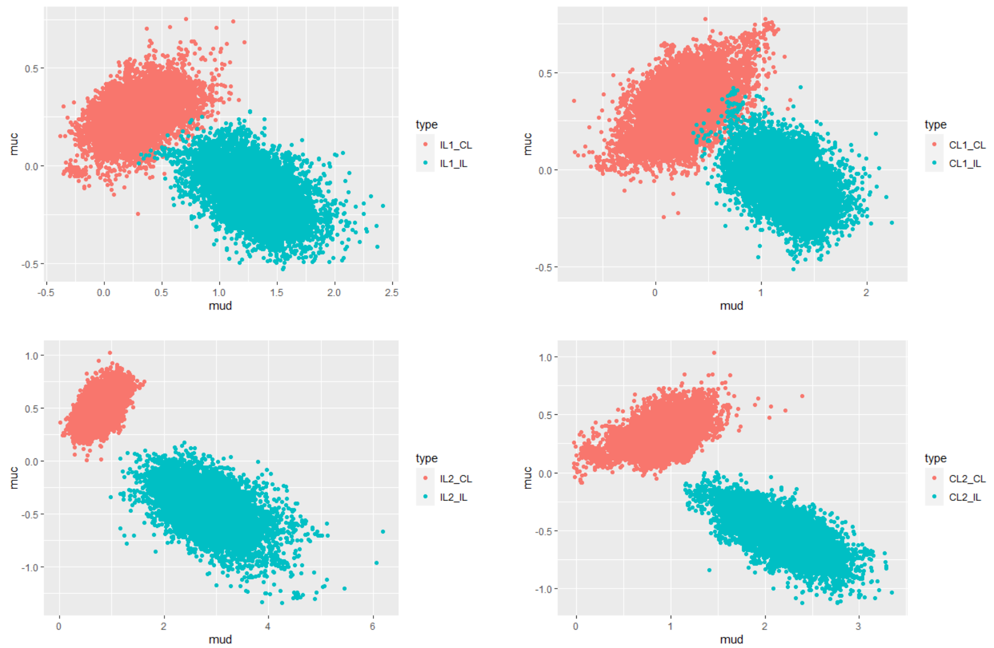


Figure 4.5: Inferred $\langle \mu_d, \mu_c \rangle$ posterior joint distributions. Left diagrams: IL group. Right diagrams: CL group. Diagrams above: the first measurements. Diagrams below: the second measurements. The shifts of the blue data clouds are apparent.

It is clear that the increase in discrimination is due to the impact of learning.

The extent of learning can be modeled with the information gain. It is therefore worth calculating the Kullbach–Leiber divergence for the discriminability distribution, which measures the improvement in learning. As it is known

$$D_{\mathrm{KL}}(\mu_d^{(2)} \parallel \mu_d^{(1)}) = \sum_{x \in X} \mu_d^{(2)}(x) \log\left(\frac{\mu_d^{(2)}(x)}{\mu_d^{(1)}(x)}\right)$$

We compute it with the R package LaplacesDemon.

group	D_{KL} in classical logic (bits)	D_{KL} in intuitionistic logic (bits)
IL	5.93	$\boxed{33.93}$
CL	$\boxed{4.35}$	17.67

The average knowledge increase of the group that studied logic with Coq is 34 bits, compared to the 4 bits of the classical logic group. Knowledge growth can also be well represented in SDT's own theoretical diagram. In the group that learned classical logic in the conventional way, the signal of classical logic inferences does not even exceed one standard deviation. In contrast, the group that studied Coq made tremendous progress in its own logic.

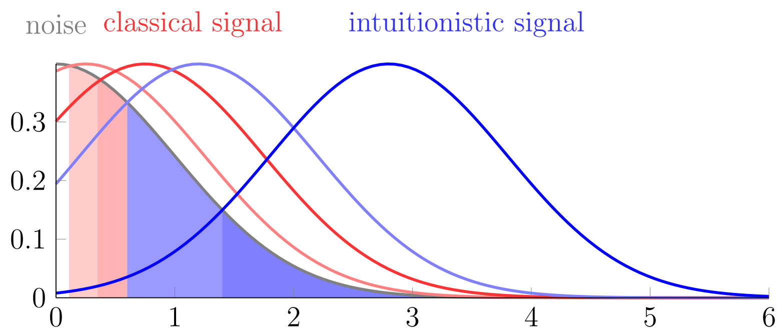


Figure 4.6: The improvement in deductive signal detection. In the CL group (red), the signal remains within one standard deviation. In the IL group (blue) the shift of the signal distribution is around 2 standard deviations.

It can even be seen from the previous figures that even the classical logical capabilities have increased in the Coq group (6 bits), although it is also true that a strong capacity increase can be observed in intuitionistic logic in the classical group as well (18 bits).

Since the distance between signal and noise is very small in the case of classical logic, it is doubtful whether there is any ability in the human brain, or at least in the case of these students' brains at their current stage of development, to recognize classical logical arguments as correct ones; in contrast to an almost perfect correlation between the responses and the evaluation of the intuitionistic logic. Followed by the learning process, a perfect correlation can be observed.

The two groups are not perfectly comparable. However, it is true that in this case, two ways of teaching are possible. First, teaching classical logic is hard to imagine with an automated theorem prover and proof management software, as they have constructivist backgrounds. So, the truth table method, advocated by Wittgenstein, remains the only possible way of teaching. However, the results show that not only is this method ineffective, but also futile, because the inner logic of the common man does not seem to follow the rules of classical logic.

When students use the computer to draw conclusions, while they are learning logic, which is more recognisable to the human brain, we get a far more effective method. Perhaps we can say that, in order to learn logic better, we should teach the computer first.

Chapter 5

Summary

The main aim of this thesis was to replicate the experiments conducted by Heit and Rotello, and to find out whether using a proof assistant software in teaching in a high school environment leads to a better understanding of logic and mathematical formalisation. The research was conducted by teaching logic to two high school groups; one was learning classical logic in a very conventional fashion, while the other was learning intuitionistic logic using a proof assistant software, called Coq. The statistical analysis of the thesis was carried out with the use of a questionnaire, which was then evaluated using Bayesian inference. The data shows that teaching classical logic the conventional way is a highly ineffective approach, whereas the use of intuitionistic logic and the implementation of computer assisted teaching methods is a much more efficient method to teaching logic. Naturally, the complete overhaul of Hungarian mathematics teaching is an overly ambitious and plain unrealistic goal, but the introduction of intuitionistic logic to the national curriculum, as well as the implementation of computer-assisted teaching methods would certainly result in an increase in the students' deductive skills, which has shown an alarming decline in the last decades.

Appendix A

Survey Questionnaire

1.

Ha Juli A, akkor Juli B is.
Juli A.

Juli B.

ValidityCL= 1

ValidityIL= 1

2.

Ha Pisti A, akkor Pisti B is.
Pisti B.

Pisti A.

ValidityCL= 0

ValidityIL= 0

3.

Ha Panni A, akkor Panni B.
Ha Panni B, akkor Panni C.
Panni A.

Panni C.

ValidityCL= 1

ValidityIL= 1

4.

Anna A, feltéve, hogy ha Anna A, akkor Anna B is.

Anna A.

ValidityCL= 1

ValidityIL= 0

5.

Ha Géza A, akkor Géza B is.

Géza vagy nem A, vagy B.

ValidityCL= 1

ValidityIL= 0

6.

Petra A, feltéve, hogy B is.

Petra A.

ValidityCL= 0

ValidityIL= 0

7.

Péter nem A.

Péter nem B.

Lehetetlen, hogy Péter A vagy B legyen.

ValidityCL= 1

ValidityIL= 1

8.

Orsi A vagy Orsi B.

Orsi A is és B is.

ValidityCL= 0

ValidityIL= 0

9.

Csabi B, feltéve, hogy Csabi A.

Csabi nem A, vagy Csabi B.

ValidityCL= 1

ValidityIL= 0

10.

Kriszti A vagy Kriszti nem B.

Kriszti B vagy Kriszti nem B.

ValidityCL= 1

ValidityIL= 0

11.

Ha Ágnes nem A, akkor Ágnes nem B.
Ágnes B.

Ágnes A.

ValidityCL= 1

ValidityIL= 0

12.

Pityu C, feltéve, hogy ha Pityu C, akkor Pityu B is.

Pityu C.

ValidityCL= 1

ValidityIL= 0

ValidityCL: classically valid or not (0/1)

ValidityIL: intuitionistically valid or not (0/1)

Appendix B

R Codes

```
options(warn = -1)
library(ggplot2)
library(ggridges)
library(ggExtra)
library(dplyr)
library(rjags)
library(coda)
library(openintro)

# DEFINE the model
hier_SDT_model <- "model{
    for (i in 1:length(H)){
    # Likelihood model
    H[i] ~ dbin(thetah[i], s[i])
    F[i] ~ dbin(thetaf[i], n[i])

    # Prior model
    thetah[i] <- phi(d[i]/2-c[i])
    thetaf[i] <- phi(-d[i]/2-c[i])
    d[i] ~ dnorm(mud, lambdad)
    c[i] ~ dnorm(muc, lambdac)
    }
```

```
    muc ~ dnorm(0, 0.001)
    mud ~ dnorm(0, 0.001)
    lambdac ~ dgamma(0.001, 0.001)
    lambdad ~ dgamma(0.001, 0.001)
    sigmad <- 1/sqrt(lambdad)
    sigmac <- 1/sqrt(lambdac)

}"

# COMPILE the model
CL1_CL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(5,3,4,4,3,5,3,3,5,4,5,3,3) ,
F= list(0,1,0,0,0,2,1,2,1,0,2,2,2),
s= list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9),
n=list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

CL1_IL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(3,2,2,3,2,3,2,2,3,2,2,2,1) ,
F= list(2,2,2,1,1,4,2,3,3,2,5,3,4),
s= list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3),
n=list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

CL2_CL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(6,5,5,5,4,6,4,4,5,6,6,2,6,4) ,
F= list(0,0,0,0,0,1,1,1,0,1,2,2,1,0),
s= list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9),
n=list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))
```

```
CL2_IL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(3,3,3,3,3,3,3,3,3,3,3,2,3,2) ,
F= list(3,2,2,2,1,4,2,2,2,4,5,2,4,2),
s= list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3),
n=list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

IL1_CL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(5,3,4,4,3,4,4,4,5,4,6,3,4,5,3,4) ,
F= list(0,0,0,0,0,2,0,1,1,2,2,1,2,2,2,2),
s= list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9),
n=list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

IL1_IL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(3,2,2,3,2,2,3,3,3,2,3,2,2,2,1,2) ,
F= list(2,1,2,1,1,4,1,2,3,4,5,2,4,5,4,4),
s= list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3),
n=list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

IL2_CL_SDT_jags <- jags.model(textConnection(hier_SDT_model),
data = list(H = list(4,5,5,3,3,4,3,3,5,4,5,4,6,3,5,4),
F= list(0,0,1,0,0,0,0,0,0,2,1,0,2,1,2,0),
s= list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9),
n=list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

IL2_IL_SDT_jags <- jags.model(textConnection(hier_SDT_model),

data = list(H = list(3,3,2,3,3,3,3,3,3,3,3,3,3,3,2,2),
F= list(1,2,4,0,0,1,0,0,2,3,3,1,5,1,5,2),
```

```
s= list(3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3,3),
n=list(9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9,9)),
inits = list(.RNG.name = "base::Wichmann-Hill", .RNG.seed = 100))

# SIMULATE the posterior

CL1_CL_sim_hf <- coda.samples(model = CL1_CL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

CL1_IL_sim_hf <- coda.samples(model = CL1_IL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

CL2_CL_sim_hf <- coda.samples(model = CL2_CL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

CL2_IL_sim_hf <- coda.samples(model = CL2_IL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

IL1_CL_sim_hf <- coda.samples(model = IL1_CL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

IL1_IL_sim_hf <- coda.samples(model = IL1_IL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

IL2_CL_sim_hf <- coda.samples(model = IL2_CL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)

IL2_IL_sim_hf <- coda.samples(model = IL2_IL_SDT_jags,
variable.names = c("muc","mud"), n.iter = 10000)


CL1_CL_stim_chains <- data.frame(CL1_CL_sim_hf[[1]],iter = 1:10000)
```

```
CL1_IL_stim_chains <- data.frame(CL1_IL_sim_hf[[1]], iter = 1:10000)

CL2_CL_stim_chains <- data.frame(CL2_CL_sim_hf[[1]], iter = 1:10000)
CL2_IL_stim_chains <- data.frame(CL2_IL_sim_hf[[1]], iter = 1:10000)

IL1_CL_stim_chains <- data.frame(IL1_CL_sim_hf[[1]], iter = 1:10000)
IL1_IL_stim_chains <- data.frame(IL1_IL_sim_hf[[1]], iter = 1:10000)

IL2_CL_stim_chains <- data.frame(IL2_CL_sim_hf[[1]], iter = 1:10000)
IL2_IL_stim_chains <- data.frame(IL2_IL_sim_hf[[1]], iter = 1:10000)

#logics compar.
CL1_logika <- rbind(CL1_CL_stim_chains, CL1_IL_stim_chains)
CL1_logika$type <- c(rep("CL1_CL",10000),rep("CL1_IL",10000))

CL2_logika <- rbind(CL2_CL_stim_chains, CL2_IL_stim_chains)
CL2_logika$type <- c(rep("CL2_CL",10000),rep("CL2_IL",10000))

IL1_logika <- rbind(IL1_CL_stim_chains, IL1_IL_stim_chains)
IL1_logika$type <- c(rep("IL1_CL",10000),rep("IL1_IL",10000))

IL2_logika <- rbind(IL2_CL_stim_chains, IL2_IL_stim_chains)
IL2_logika$type <- c(rep("IL2_CL",10000),rep("IL2_IL",10000))

#group comp

ALL1_stim <- rbind(CL1_CL_stim_chains,IL1_IL_stim_chains)
ALL1_stim$type <- c(rep("CL1",10000),rep("IL1",10000))

ALL2_stim <- rbind(CL2_CL_stim_chains,IL2_IL_stim_chains)
ALL2_stim$type <- c(rep("CL2",10000),rep("IL2",10000))
```

```
#improvement

CL_CL_fejl <- rbind(CL1_CL_stim_chains, CL2_CL_stim_chains)
CL_CL_fejl$type <- c(rep("CL1_CL",10000),rep("CL2_CL",10000))

CL_IL_fejl <- rbind(CL1_IL_stim_chains, CL2_IL_stim_chains)
CL_IL_fejl$type <- c(rep("CL1_IL",10000),rep("CL2_IL",10000))

IL_CL_fejl <- rbind(IL1_CL_stim_chains, IL2_CL_stim_chains)
IL_CL_fejl$type <- c(rep("IL1_CL",10000),rep("IL2_CL",10000))

IL_IL_fejl <- rbind(IL1_IL_stim_chains, IL2_IL_stim_chains)
IL_IL_fejl$type <- c(rep("IL1_IL",10000),rep("IL2_IL",10000))

##KLD##
#KLDILIL

N1 <- coda.samples(model = IL1_IL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x1<-sample(N1, size=1, replace = T)
unx1<-unlist(x1)
mean(unx1)
sd(unx1)

N2 <- coda.samples(model = IL2_IL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x2<-sample(N2, size=1, replace = T)
unx2<-unlist(x2)
mean(unx2)
sd(unx2)

z_scores <- seq(-1, 5, by = .001)
```

```
norm1<-dnorm(z_scores, mean(unx1), sd(unx1))
norm2<-dnorm(z_scores, mean(unx2), sd(unx2))
KL<-KLD(norm1,norm2, base=2)
KL

#KLDILCL

N1 <- coda.samples(model = IL1_CL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x1<-sample(N1, size=1, replace = T)
unx1<-unlist(x1)
mean(unx1)
sd(unx1)

N2 <- coda.samples(model = IL2_CL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x2<-sample(N2, size=1, replace = T)
unx2<-unlist(x2)
mean(unx2)
sd(unx2)

z_scores <- seq(-1, 5, by = .001)

norm1<-dnorm(z_scores, mean(unx1), sd(unx1))
norm2<-dnorm(z_scores, mean(unx2), sd(unx2))
KL<-KLD(norm1,norm2, base=2)
KL

#KLDCLIL

N1 <- coda.samples(model = CL1_IL_SDT_jags, variable.names = c("mud"),
```

```
n.iter = 10000)
x1<-sample(N1, size=1, replace = T)
unx1<-unlist(x1)
mean(unx1)
sd(unx1)

N2 <- coda.samples(model = CL2_IL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x2<-sample(N2, size=1, replace = T)
unx2<-unlist(x2)
mean(unx2)
sd(unx2)

z_scores <- seq(-1, 5, by = .001)

norm1<-dnorm(z_scores, mean(unx1), sd(unx1))
norm2<-dnorm(z_scores, mean(unx2), sd(unx2))
KL<-KLD(norm1,norm2, base=2)
KL

#KLDCLCL

N1 <- coda.samples(model = CL1_CL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x1<-sample(N1, size=1, replace = T)
unx1<-unlist(x1)
mean(unx1)
sd(unx1)

N2 <- coda.samples(model = CL2_CL_SDT_jags, variable.names = c("mud"),
n.iter = 10000)
x2<-sample(N2, size=1, replace = T)
```

```
unx2<-unlist(x2)
mean(unx2)
sd(unx2)

z_scores <- seq(-1, 5, by = .001)

norm1<-dnorm(z_scores, mean(unx1), sd(unx1))
norm2<-dnorm(z_scores, mean(unx2), sd(unx2))
KL<-KLD(norm1,norm2, base=2)
KL
```

Bibliography

[1] Aristotle. *Aristotle's Prior Analytics*. Oxford University Press, 2009.

[2] Gábor Dancs. "A matematikus abszolút pedagógusról, Varga Tamásról". In: *Abszolút pedagógusok. Új szempontok a XX. századi értelmiségtörténet kutatásához*. Ed. by László Trencsényi Endre Kiss and Árpád Hudra. Budapest, HU: Kaleidoscope könyvek (5). LÉTRA Alapítvány - Magyar Pedagógiai Társaság, 2021, pp. 238–252. ISBN: 978-615-6275-01-1. URL: http://real-eod.mtak.hu/9617/1/238-252_abszolut_pedagogusok_2021.pdf.

[3] György Pólya. *A problémamegoldás iskolája I.* Reprint. TypoTeX, 2010. ISBN: 978-963-2791-25-8.

[4] Janos Gyori et al. "The Traditions and Contemporary Characteristics of Mathematics Education in Hungary in the Post-Socialist Era". In: May 2020, pp. 101–104. ISBN: 978-3-030-38743-3. DOI: 10.1007/978-3-030-38744-0_3.

[5] Hoyles C. "A culture of proving in school mathematics?" In: *Information and Communications Technologies in School Mathematics.* Ed. by Johnson D.C. Tinsley D. Boston, MA: Springer, 1998, pp. 169–182. URL: https://link.springer.com/content/pdf/10.1007/978-0-387-35287-9_21.pdf.

[6] Péter Bereczky et al. "Interactive Teaching of Programming Language Theory with a Proof Assistant". In: 2020.

[7] Sebastian Böhne and Christoph Kreitz. "Learning how to Prove: From the Coq Proof Assistant to Textbook Style". In: *CoRR* abs/1803.01466 (2017), pp. 1–18.

[8] Lance J. Rips. "The Psychology of Proof: Deductive Reasoning in Human Thinking". In: 1994.

[9] *The Cambridge Handbook of Computational Psychology*. Cambridge Handbooks in Psychology. Cambridge University Press, 2008. DOI: 10.1017/CBO9780511816772.

[10] B. Skyrms. *Choice and chance: An introduction to inductive logic*. Fourth edition. Belmont, CA, USA: Wadsworth, 2000, pp. 17–23. ISBN: 978-053-4557-37-9.

[11] Evan Heit and Caren M. Rotello. "Are There Two Kinds of Reasoning". In: *Proceedings of the Annual Meeting of the Cognitive Science Society* 27 (2005). URL: https://escholarship.org/uc/item/58k5m8k3.

[12] L. J. Rips. "Two Kinds of Reasoning". In: *Psychological Science* 12.2 (2001), pp. 129–134. DOI: https://doi.org/10.1111/1467-9280.00322.

[13] Michael Lee. "Three case studies in the Bayesian analysis of cognitive models". In: *Psychonomic bulletin review* 15 (Mar. 2008), pp. 1–15. DOI: 10.3758/PBR.15.1.1.

[14] Wagenmakers E. Lee M. *Bayesian Cognitive Modeling: A Practical Course*. Fourth edition. Cambridge, UK: Cambridge University Press, 2014, pp. 156–164. ISBN: 978-113-9087-75-9. DOI: 10.1017/CBO9781139087759.

[15] Lu J. Rouder JN. "An introduction to Bayesian hierarchical models with an application in the theory of signal detection". In: *Psychonomic Bulletin Review* 12 (2005), pp. 573–604. DOI: 10.3758/bf03196750.

[16] R.Harald Baayen, Fiona J. Tweedie, and Robert Schreuder. "The Subjects as a Simple Random Effect Fallacy: Subject Variability and Morphological Family Effects in the Mental Lexicon". In: *Brain and Language* 81.1 (2002), pp. 55–65. ISSN: 0093-934X. DOI: https://doi.org/10.1006/brln.2001.2506. URL: https://www.sciencedirect.com/science/article/pii/S0093934X01925064.

[17] W.R. Gilks, S. Richardson, and D. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall/CRC Interdisciplinary Statistics. Taylor & Francis, 1995. ISBN: 9780412055515. DOI: 10.1201/b14835. URL: http://books.google.com/books?id=TRXrMWY_i2IC.

[18] Richard J Herrnstein. *The bell curve : intelligence and class structure in American life*. The Free Press, 1994, p. 44. ISBN: 0029146739.

[19] John L. Bell, David DeVidi, and Graham Solomon. *Logical Options: An Introduction to Classical and Alternative Logics*. Broadview Press, 2001, pp. 1–47.

[20] Stewart Shapiro and Teresa Kouri Kissel. “Classical Logic”. In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Spring 2021. Metaphysics Research Lab, Stanford University, 2021.