Lisbon School of Economics & Management
Universidade de Lisboa

MASTER

ECONOMICS

MASTER'S FINAL WORK

DISSERTATION

CAN COMPLETE INFORMATION ON PAST COHORT PERFORMANCE BREAK STUDENTS' OVERCONFIDENCE? AN EXPERIMENTAL APPROACH USING A REAL-EFFORT TASK

MARTA MORGADO ROSA

JULY – 2024

ISEG
Lisbon School of Economics & Management
Universidade de Lisboa

Master
Economics

Master's Final Work
Dissertation

Can Complete Information on Past Cohort Performance Break Students' Overconfidence? An Experimental Approach Using a Real-Effort Task

Marta Morgado Rosa

Supervision:
Maria Joana Dantas Vaz Pais
Matthijs Oosterveen

July – 2024

GLOSSARY

IQR – Interquartile Range.

GPA – Grade Average Point.

ISEG – Lisbon School of Economics and Management.

MSE – Mean Squared Error.

ORSEE – Online Recruitment System for Economic Experiments.

RCTs – Randomised Controlled Trials.

RMSE – Root Mean Squared Error.

SOJs – Second-order Judgements.

XLAB – Behavioural Research Lab.

ABSTRACT, KEYWORDS AND JEL CODES

This work investigates the influence of information about the performance of previous cohorts on the accuracy of predictions. Overconfidence is a cognitive bias that can lead students to underestimate the effort required to obtain the desired grades, affecting their academic success. To address this problem, I carried out a lab experiment using a between-subjects design, dividing the participants into four treatments. When predicting, the control groups (T1 and T2) had no information, while the treatment groups (T3 and T4) had information that included visual data on the distribution of performance from previous cohorts. The experiment followed a methodology derived from Abeler et al. (2011), using a real-effort task involving counting zeros in tables to assess performance accuracy. Participants' mindsets and self-esteem were also measured through specific questions, allowing categorisation into deliberative or implemental mindsets and levels of self-esteem through the Rosenberg Self-Esteem Scale.

The main conclusion of the study was that providing information immediately reduces prediction errors, with the treatment group showing a median prediction error of 0. This result emphasises the potential benefits of such interventions in educational contexts to improve students' grade predictions and enhance their academic performance. The final model, with 16 predictors, explained 86.6% of the variability in under/overconfidence, revealing the importance of actual performance and mindset. A Bayesian network model also indicated that actual performance on the task was crucial, although residual patterns suggested areas for further investigation.

KEYWORDS: Accuracy; Forecasts; Overconfidence; Predictions; Students.

JEL CODES: A20; C11; C91; D83; D84; D90; D91.

TABLE OF CONTENTS

Glossary i

Abstract, Keywords and JEL Codes ii

Table of Contents iii

Table of Figures v

Acknowledgments viii

1. Introduction 1

2. Literature Review 3

3. Experimental Design and Implementation 7

3.1. Treatments 7

3.2. Variables 9

3.3. Procedures 10

Pre-test 10

Experiment 11

3.4. Payment Scheme 13

3.5. Procedures 14

3.6. Ethical Considerations 14

3.7. Hypotheses and Analytical Framework 15

4. Results 17

4.1. Descriptive statistics 17

4.2. Immediate effects of the intervention 20

4.3. Absolute Under/Overconfidence 21

4.4. Relative Under/overconfidence 24

4.5. Learning 25

4.6. Mindset 28

4.7. Self-esteem 29

4.8. Regression Models 31

4.9. Bayesian Network 39

5. Conclusion 44

6. References 47

Appendices 53

Appendix A – Wilcoxon Rank-Sum test results for A[n] and E[n] 53

Appendix B – Pre-test Link and Instructions 53

Qualtrics link 53

Detailed Instructions 53

Appendix C – Experimental Links 55

Appendix D – Photo of the Laboratory at XLAB 56

Appendix E – Ethical Approval 57

Appendix F – Expected and Actual Performance Graphs 58

Appendix G – Predictors of avg_delta with the interaction term 65

Appendix H – Experimental databases and codes 66

TABLE OF FIGURES

Figure 1 – Sequence of events during the experiment.. 13

Figure 2 – Variable payment scheme table. 14

Figure 3 – Total earnings considering all rounds, per condition. 19

Figure 4 – Absolute Δ_1 for the control and treatment groups. 21

Figure 5 – Absolute Δ per round in T1 and T3. 23

Figure 6 – Absolute Δ per round in T2 and T4. 23

Figure 7 – On the left, the average absolute Δ (for all rounds) by grade, for T1 and T3. On the right, the average absolute Δ (for all rounds) by grade, for T2 and T4. 24

Figure 8 – Relative Δ per round in T1 and T3. 25

Figure 9 – Relative Δ per round in T2 and T4. 25

Figure 10 – On the left, the average absolute Δ difference between rounds, for T1 and T3. On the right, the average absolute Δ difference between rounds, for T2 and T4..... 27

Figure 11 – On the left, the average absolute Δ depending on the mindset, for T1 and T3. On the left, the average absolute Δ depending on the mindset, for T2 and T4. 29

Figure 12 – The correlation between self-esteem (SE) and average absolute Δ in each of the conditions. 31

Figure 13 – Correlation matrix of variables represented in the regression model 2, using a color-coded and size-scaled bubble plot. 33

Figure 14 – Residuals vs Fitted Plot from the regression model 4. 36

Figure 15 – Normal Q-Q Plot from the regression model 4. 37

Figure 16 – Standardized residuals against the leverage of each observation in the regression model 4. 38

Figure 17 – Relative importance of the variables in the regression model 4. 38

Figure 18 – Custom Bayesian network structure for predicting avg_delta. 40

Figure 19 – Relationship between the observed values and the predicted values generated by the Bayesian Network. 41

Figure 20 – Residuals vs Predicted Plot from the Bayesian Network. 42

Figure 21 – Relative importance of the variables in the Bayesian network. 43

Figure 22 – Correlation matrix for the Bayesian network. 43

Figure 23 – Density plots for the Bayesian network. 44

Figure 24 – On the left, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T1. On the right, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T3. 58

Figure 25 – On the left, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T2. On the right, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round in, T4. 59

Figure 26 – On the left, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T1. On the right, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T3. 60

Figure 27 – On the left, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T2. On the right, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T4. 61

Figure 28 – On the left, the relationship between the expected percentile, e[n], converted into quartiles, relative to peers vs the average actual number of tables correctly counted, A[n], before the task was known, between T1 and T3. On the right, the relationship between the expected percentile, e[n], converted into quartiles, relative to peers vs the average actual number of tables correctly counted, A[n], before the task was known, between T2 and T4. 62

Figure 29 – On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs the expected percentile relative to peers, e[n], per round, in T1. On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs expected percentile relative to peers, e[n], per round, in T3. 63

Figure 30 – On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs the expected percentile relative to peers, e[n], per round, in T2. On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs expected percentile relative to peers, e[n], per round, in T4. 64

ACKNOWLEDGMENTS

Before anything else, I would like to express my deepest gratitude to my supervisors, Professors Joana Pais and Matthijs Oosterveen. Knowing that I had their support during this period was a good safety net. They allowed me to work entirely autonomously and guided me in finding my next adventure. A special mention to Professor Joana, who offered me a diverse range of experiences, followed the experimental phase of my dissertation and, most importantly, encouraged me to embrace a change in my academic path. It was the push I needed to not be afraid to challenge myself!

At the same time, during one of my classes at FCH Católica, I met Professor Cláudia Simão. Although I initially fell in love with Experimental Economics at ISEG, which was why I decided to pursue my master's degree here, I was truly disappointed with the potential of this field when it was time to create my research question. Without Professor Cláudia's influence, I would likely have ended up miserably writing about macroeconomic models. I am grateful to her for giving me back the smile I used to wear when I discovered new games or concepts when learning about Experimental Methods.

Attending a Master's degree has its ups and downs, and without my best friend, Margarida Almeida, who was there at every important stage of my academic life, these years would have been much more monotonous. I would also like to thank my colleague, now friend, Regina Pereira, whom I met during this program. In addition to our vacations together, she made our study moments enjoyable and was always willing to help me during the XLAB sessions, ensuring the anonymity of the participants in my study. Equally important, I must emphasise the endless companionship of my friend Francisca Martins during the days I was writing this document.

Finally, this journey would not have been possible without my parents, who have financed my education, even when they do not agree with the decisions I make. They give me the freedom to choose the career I am now beginning to build.

¡Hasta luego!

CAN COMPLETE INFORMATION ON PAST COHORT PERFORMANCE BREAK STUDENTS' OVERCONFIDENCE? AN EXPERIMENTAL APPROACH USING A REAL-EFFORT TASK

By Marta Morgado Rosa

1. INTRODUCTION

It is estimated that the average adult makes around 35,000 decisions a day. Some are quick and straightforward, while others require considerable cognitive effort. Individuals face particular difficulties in evaluating decisions where the outcomes are uncertain, which brings expectations and beliefs into the equation (Didier Demazière, 2024). These expectations and beliefs are often irrational, as mentioned in Gric et al. (2022), or based on incomplete information, as explained in Han (2020), making the correct choice even harder to make.

In the context of the economics of education, this is particularly worrying given Portugal's spending on educational institutions. According to the OECD, in 2023, Portugal spent 5.1% of its GDP on education, from primary to university level (OECD, 2023). Thus, from an economic perspective, improving students' ability to accurately predict their grades makes sense, as it can lead to better academic results, which in turn can increase the overall efficiency of the education system.

To explore this issue, it is important to understand the daily decision-making processes that students go through. Each choice, from selecting a course unit to predicting final grades, involves varying degrees of intellectual effort. This is where the complexity of human cognition comes into play. For instance, predicting the final grade for a course, if the assessment is continuous, involves prior forecasts for several test grades. Especially when done for the first time, the difficulty in accurately predicting results is higher (Subramaniam, 2022), meaning that beliefs and expectations may not remain stable during the semester (Snyder et al., 2018).

One of the main psychological phenomena relevant to this discussion is overconfidence. In Kahneman (2011) overconfidence was labelled "the most significant of the cognitive biases". Overconfidence occurs whenever individuals overestimate their abilities or the accuracy of their predictions. Conversely, underconfidence manifests itself when prediction values are lower than actual performance.

In the context of academic performance, overconfidence can lead students to underestimate the amount of effort needed to achieve the desired grades. This mismatch between expectations and reality can have harmful effects on students' academic success and general well-being. Some might argue that professors' predictions can serve as a reliable starting point. However, these often penalises certain groups, such as top-performing students from disadvantaged backgrounds (Wyness et al., 2022). Even so, first-year expectations are a good indicator of retention in the second semester (Acee et al., 2020).

Although students do not follow the model of the rational agent (*homo economicus*) and sometimes display irrational behaviours, according to learning theory, students, like other human beings, receive, process, and retain new knowledge (Conner, 2022; Pennings et al., 2019). This study focuses on students' expectations of their grades, a particularly difficult task when entering a new discipline. These expectations have consequences beyond the immediate, influencing the study intentions needed throughout the semester to achieve the desired grades. Thus, cognitive biases represent an opportunity for policymakers to design public policies to improve academic performance through teaching methods.

From a complementary perspective, it is possible to shape behavior by being aware of its biases (Sagar et al., 2019), and, for that, psychology also plays an important role in understanding human behavior. Mindset theory distinguishes between an implemental mindset, centered on planning goal-oriented actions, and a deliberative mindset, known for its consideration of the pros and cons behind the adoption of a particular goal (Gollwitzer et al., 1990). A more deliberate decision-making process, centered on real capabilities and not just desired objectives, is associated with more accurate scenarios (Keller & Gollwitzer, 2017). From another view, uncertainty can be an opportunity to demonstrate high self-esteem, as found in Yang et al. (2019), but it is also associated with risky attitudes, where the end goal often takes precedence over the means, leaving open the results that could come from its combination with predictive skills.

In this study, I will conduct a lab experiment to take advantage of the anchoring bias – in which individuals rely too heavily on the first information they encounter while making decisions. The main objective is to test a virtually cost-free intervention that can contribute to the academic success of university students. Improving students' ability to make accurate predictions about their academic performance can lead to more targeted and efficient study strategies. The findings of this study can serve as a basis for designing other low-cost interventions that lecturers can implement in the classroom context to potentially improve class averages.

The work is organized as follows. The current state of the literature is presented in the Literature Review. The design, procedures, and hypotheses are described in the Experimental Design and Implementation section. The results of the hypotheses to test overconfidence and its relationship with mindset and self-esteem are presented in the Results section. To give a predictive perspective, I have also added regression models and a Bayesian Network to this section. Finally, the Conclusions section offers some closing remarks, limitations, and suggestions for future research.

2. LITERATURE REVIEW

The literature on overconfidence in academic contexts highlights several key factors that influence students' expectations about performance. These factors include the provision of information (Guskey, 2022), the effectiveness of feedback mechanisms (Erat et al., 2020; Saenz et al., 2019), the role of monetary incentives (Ruthig & Kroke, 2024), and individual characteristics (Geraci et al., 2022; Hamann et al., 2020; Silva et al., 2021; Vignery, 2022).

Several studies have shown that students do not rationalise their expectations (Hossain & Tsigaris, 2012; Magnus & Peresetsky, 2018). At first, there is an estimation error that improves as there is more information. In a field study by Wright & Arora (2022), students were provided with instructor-specific information about the past distribution of grades in their Principles of Macroeconomics course. The intervention led to a 10-percentage point (pp) increase in the probability of passing. Students with high prior expectations were the ones who benefited most from the treatment, as the information provided helped them adjust their expectations downward to more closely fit with the actual grade distribution. This adjustment in expectations resulted in more realistic self-assessments and improved their chances of academic success, two outcomes crucial for long-term academic and personal development.

In the laboratory context, the results of Abeler et al. (2011) were already in line with models of expectation-based reference-dependent preferences. To do this, they used a real-effort task, which consisted of counting the number of zeros in tables. The experiment involved two main treatments with different fixed payment amounts. Participants in the high fixed payment treatment of 7 euros worked significantly longer than those in the low fixed payment treatment of 3 euros, demonstrating a clear treatment effect on effort provision. On average, participants in the high fixed payment treatment stopped working after accumulating 9.22 euros, whereas those in the low fixed payment treatment stopped at 7.37 euros.

Similarly, it has been reported that giving early and frequent feedback helps students to calibrate their grade expectations more realistically (Koenka, 2020). This is because, as Armstrong & MacKenzie (2017) argue in support of self-regulated learning theory, students' study is the result of a cyclical pattern adjusted according to the feedback received. The study surveyed 278 students in a first-year undergraduate business course, collecting data on their grade expectations, actual grades, and studying intentions during two moments of the semester. The findings revealed that students increased their studying efforts if their actual grades were lower than their original or updated goals. Conversely, the difference between students' subjective grade goals and their objectively forecasted final grades did not significantly influence their studying intentions. This suggests that students react more effectively to immediate performance feedback than to predictive forecasts.

However, according to Nederhand et al. (2020), this process is not significantly dependent on the level of reflection support provided. Their longitudinal quasi-experimental study conducted in a secondary school found that simply asking students to estimate their grades and providing regular feedback on these estimates reduced the forecasting error, regardless of whether students were given additional reflection support. On the other hand, it was observed that students who were part of the reflection group showed a significant increase in their second-order judgments (SOJs), which are measures of confidence in their performance estimates, compared to other groups. This reveals the importance of regular self-assessment practices and the potential benefits of incorporating reflection to boost student confidence, ultimately leading to better academic performance.

Nonetheless, not all groups of students react in the same way to interventions designed to reduce overconfidence. Tirso et al. (2019) asked students to predict their grades, percentile ranks and class averages before taking exams. The findings revealed that top-performing (or "A") students accurately predicted their own grades but underestimated their relative standing compared to peers, indicating underconfidence. Contrary to the false consensus hypothesis – which argues that "A" students overestimate the abilities of their peers because they assume the tasks are easier for everyone – top performers did not overestimate more the class average compared to bottom performers ("D" students). Further, the study found that while low performers consistently showed overconfidence in their grade predictions and percentile ranks, top performers, despite receiving feedback, were persistently underconfident in their relative positioning.

The above conclusions were built on the work done by Ehrlinger et al. (2008) in which five studies explored how the perceived simplicity of tasks affects performance predictions. They reported that when tasks are perceived as easy, bottom performers are more likely to overestimate their performance. This is because the simplicity of the task leads them to believe that their performance is better than it actually is. On the opposite side, top performers might underestimate their relative performance on simple tasks because they assume that if the task is easy for them, it must be easy for everyone else as well. This perceived simplicity contributes to their underconfidence in relative performance despite their accurate self-assessment in absolute terms. Thus, task difficulty is a variable to consider when defining the accuracy of the self-assessment of low and high performers.

In relation to monetary incentives, Sabater-Grande et al. (2022) conducted a study in a classroom setting involving post-exam predictions, with randomized groups receiving monetary incentives and others without. It was found that students tend to overestimate their performance. Yet, there is evidence of the Dunning-Kruger effect, with bottom-performing students showing a greater overestimation. Introducing monetary incentives significantly reduced students' overestimation when setting their own target grades and in the post-diction grades immediately after the exams. This reduction was not due to more accurate guesses but rather due to better academic performance, suggesting that incorporating monetary rewards into educational policies may help students to enhance their academic outcomes.

By implementing an extra credit scheme, Caplan et al. (2017) rewarded students with extra credit when accurately predicting the number of multiple-choice questions they would answer correctly on an upcoming exam. The incentive scheme, inspired by a bonus model used by IBM in Brazil, not only reduced instances of extreme forecast errors but also improved the overall forecast accuracy across different student groups. For top-performing students, the scheme mitigated extreme optimism among "B" students and reduced instances of extreme pessimism for "A" students. Still, the study also noted that while incentives can improve accuracy for some students, they may also induce risk-seeking behaviour in others, particularly those with lower performance levels, resulting in larger forecast errors for this group.

Building on the idea that decision-making is highly influenced by individual characteristics, research exploring different cognitive orientations is fundamental in understanding the mental frameworks that, in this case, lead individuals to attain a goal. For that, let's first define the two concepts involved in this dual-process model notion – implemental and deliberative mindset. The first one occurs whenever individuals start planning on how to attain a specific goal. For that, they purely rely on goal-related information, meaning that they are closed-minded and biased (Bayer & Gollwitzer, 2005). Alternatively, in the deliberative mindset, the pros and cons of the desirability of the goal are evaluated as well as its feasibility. Thus, there is a realistic consideration of the potentialities of the self, accompanied by an open-mindedness to all sources of information (Brandstätter et al., 2015).

Li et al. (2018) examined the influence of deliberative and implemental mindsets on decision-making, particularly in scenarios that require Bayesian updating. Participants with a deliberative mindset were more prone to reinforcement and comprehension errors, as a result of a tendency to think too much and be distracted by irrelevant information. This mindset did not support systematic, goal-centred processing, so it did not always translate into better decision-making results. On the other hand, participants with an implemental mindset were more skilled at incorporating new information with prior beliefs, adhering closely to Bayes' rule and consequently making more accurate decisions. This mindset reduced reinforcement errors, since individuals typically follow a heuristic of about past actions well: if successful, they decide to repeat, if unsuccessful they decide to switch to avoid past failures. In this way, information processing becomes more thorough and concentrated. This shows a connection between mindset and efficient, rational decision-making, especially in complex and/or uncertain economic scenarios.

Other connections that should be taken into consideration include gender and ethnic differences (Khattab et al., 2021), risk behaviours and perceptions (Keller & Gollwitzer, 2017), traces of narcissism (O'Reilly & Hall, 2021), among others. As not all variables can be analysed at the same time, I would like to highlight self-esteem because, as described in Hügelschäfer & Achtziger (2014), implemental mindset participants have higher levels of self-esteem, which in turn makes them less vulnerable to anchors that may disturb their focus. Thus, this is the last element of the triangulation that I will use in my study.

Most of the studies mentioned before are either lab, with cognitive tests, or field experiments or survey-based. While various social situations and behaviours have been studied using real-effort tasks (Chapkovski & Kujansuu, 2019; Jiménez-Jiménez et al., 2023; Rodrigo-González et al., 2021), there is still a gap in research on overconfidence among university students, especially regarding their grade predictions throughout the semester. The research questions about the impact of giving information about past cohort performance are as follows: Is the effect immediate? Are students more accurate in their prior predictions of their own performance? Are they able to position themselves more correctly in relation to their peers? Do students learn to be more accurate faster? Is there a specific mindset associated with more accurate predictions? Can the relationship between self-esteem and overconfidence be weakened?

3. EXPERIMENTAL DESIGN AND IMPLEMENTATION

3.1. Treatments

In this study, a between-subject design was used to assess the impact of information on past cohort performance in making university students' forecasts more accurate. As such, participants were divided into four treatments, henceforth mentioned as T1, T2, T3 and T4. Each of them comprised three rounds in which three similar tasks had to be completed. The control groups are T1 and T2, while T3 and T4 constitute the treatment groups. Before each round, participants had to forecast their performance in the following task. At that moment, the control groups had no information about the past performance of others, while those in the treatment group, were presented with information about the performance of a past cohort who did the same task as them.

To guarantee that any learning between rounds was not the result of systematic differences in difficulty levels, which were necessary to keep the answers unpredictable, the order of rounds 2 and 3 was switched between T1 (T3) and T2 (T4) within the control (treatment) groups. Specifically, for T1 and T3 the last two rounds had one sequence, while for T2 and T4 the sequence was the opposite, as presented in Table I. This detail was made so that the order of rounds would not be a confounder, i.e. to exclude learning caused by a reduction of the level of difficulty of tasks throughout the rounds. However, to allow for a clean comparative analysis between groups and to assess the immediate effect of the intervention, round 1 was kept in the same position for all subjects.

TABLE I

DIFFERENCES BETWEEN TREATMENTS

	T1	T2	T3	T4
Information	No	No	Yes	Yes
Forecasts	Before each round	Before each round	Before each round	Before each round
Feedback	After each round	After each round	After each round	After each round
Order of rounds	1→2→3	1→3→2	1→2→3	1→3→2

Initially, I tried to join the two conditions in the control group and the two in the treatment group. However, I found significant differences in one of the variables that contribute to the payoff for T1 and T2 in Round 3 ($p = 0.000 < 0.001$) and for T3 and T4 in Round 2 ($p = 0.001 < 0.01$)[1]. These significant differences indicate that the groups are not comparable and should not be joined. To avoid compromising the internal validity, as

[1] See Appendix A for the Wilcoxon Rank-Sum test results for the relevant variables.

there was the possibility that students who had already participated could share some details of the experiment with their friends, sessions for T1 and T2 were conducted in the first two days, and the sessions for T3 and T4 took place on the following days.

3.2. Variables

Table II shows the factors influencing the analysis of this experiment. The independent variable, condition (COND), is a nominal variable that defines the experimental condition assigned to participants, that can take one of four classifications (T1, T2, T3 and T4). In each round, I measured some dependent variables to determine students' own expectations and their evaluation of their performance relative to others, denoted as E[n] and e[n] respectively, which were what they were asked to forecast. Additionally, I recorded their actual performance (A[n]) and their real quartile positioning (a[n]), which was converted into a letter grade (ranging from D to A) due to the impracticality of using a continuous scale. This is because on a continuous scale, applied on small datasets with limited score variations, such as this one, where the number of correct answers per round could only range from 0 to 5, large clusters of participants could fall into the same percentile or range. As a consequence, the scale could possibly not provide any additional useful information beyond what a simpler categorization (like letter grades) could offer. Several control variables related to the dual-process theory (MIND), self-esteem confidence levels (SE), and demographics were also collected, as detailed below.

TABLE II

VARIABLES USED FOR THE ANALYSIS

Acronym	Explanation	Assignment
COND	Condition to which the participant is assigned	Ordinal variable: 1 if T1, 2 if T2, 3 if T3, 4 if T4
Expectations variables		
E[n]	Expected number of tables correctly counted in the nth round	Interval variable: 0, 1, …, 5 tables
e[n]	Expected percentile relative to the peers in the nth round	Interval variable: 1, 2, …, 99 percentile
Actual variables		
A[n]	Actual number of tables correctly counted in the nth round	Interval variable: 0, 1, …, 5 tables
a[n]	Actual quartile relative to their peers in the nth round	Interval variable: D, C, B, A grade
Other control variables		
MIND	Mindset state	Binary variable: 0 if implemental, 1 if deliberative
SE	Self-esteem scale positioning	Interval variable: 10, 11, …, 40
AGE	Age	Ordinal variable: 18, 19, …, >30
GEN	Gender identification	Nominal variable: 1 if female, 2 if male, 3 if non-binary, 4 if other
CYL	Current university cycle	Ordinal variable: 1 if bachelor's degree, 2 if master's degree, 3 if PhD degree
YEAR	Course year	Ordinal variable: 1 if 1st, 2 if 2nd, 3 if 3rd, 4 if other[1]
GPA	Current GPA	Interval variable: 10, 11, …, 20

[1] All participants who chose "other" in the variable "YEAR" reported the number 4.
n ∈ {1, 2, 3}

3.3. Procedures

Pre-test

Before implementing the experiment, a pre-test was conducted to ensure the reliability of the task design. In order to be able to distinguish the top from the bottom performers, the real-effort task, which according to Charness et al. (2018) is a way of measuring the level of effort exerted by the participants, needed a level of difficulty that allowed for clear differentiation in performance. Based on these results, two adjustments were made before finalizing the design of the experiment. Firstly, because the level of the tasks was too difficult and most of the participants were not even able to finish the tasks within the time limit that existed, and then the opposite occurred, leading to a performance in which it was difficult to identify top performers.

The pre-test took place at ISEG, where participants were recruited through two methods: by inviting students attending bachelor classes and by directly approaching them on campus. Participation was voluntary and the pre-test had a maximum duration of 8 minutes. Before the end of the pre-test, participants were required to create unique IDs to identify themselves at the time of payment. In total, 33 observations were collected.

Unlike the main study, which evaluated the accuracy of forecasts, the pre-test aimed to test the task. Therefore, students were compensated with a specific and easier-to-understand payment scheme[2]. Moreover, because the environment was less controlled than a laboratory, questions designed to control for confounding variables were left out. A lottery was conducted two days after the end of the pre-test, allowing five participants from this pool to win up to 10 euros in Gift Cards. The winning IDs were announced through XLAB's Instagram page.

Experiment

When arriving at the laboratory, each subject was randomly assigned to one of the conditions to secure an equal number of participants in each treatment. To minimize potential biases or peer effects (Basse et al., 2024), all participants were seated in identical places. Each session included participants from either conditions T1 and T2, or T3 and T4, alternately.

The experiment follows a methodology derived from Abeler et al. (2011), applying a similar real-effort task, counting the number of zeros in tables[3], albeit with different objectives. This task was selected because it does not require previous specific knowledge and is performed individually. The tables designed using Jupyter Notebook, a web-based

[2] The link and the detailed instructions, including the payment scheme, can be found in Appendix B.

[3] Appendix C contains the links to the experimental conditions.

interactive computing platform, contained 150 randomly placed zeros and ones constrained by the following condition:

$(1)\ 33 < \text{number of zeros} < 74.$

In the control groups, after reading the instructions and viewing an example table to ensure a solid understanding, participants began round 1. Firstly, they encountered a variable payment table, which was a matrix used to determine the part of the payment that was not fixed and was dependent on their behaviour, and set their predictions about the number of tables they expected to count correctly (E[1]) and their performance relative to others (e[1]). Only then were they allowed to view the 5 tables that existed per round. Secondly, they counted the number of zeros in as many tables as possible correctly within 2 minutes. After this time expired or all tables were counted, the study automatically progressed to a screen where participants received feedback on the accuracy of their predictions, allowing SOJs. As such they had the possibility of adjusting their estimates' errors in subsequent rounds. This process was repeated in rounds 2 and 3, to mimic a university course with continuous assessment at 3 different points in time.

Following the task stage, participants answered three dichotomous questions related to their performance, each indicating a propensity for being in either an implemental or deliberative mindset while completing the tasks. To determine the predominance of their mindset, the mode of their responses to these questions was then used, allowing me to classify them as being in an implemental or deliberative mindset.

Then, they completed a 10-item questionnaire (Rosenberg, 1965) using a four-point Likert scale ranging from "Strongly Disagree" to "Strongly Agree". These questions assess participants' feelings of self-worth and self-acceptance throughout their life (and not just concerning this experiment), including positive statements like "On the whole, I am satisfied with myself." as well as negative statements like "I certainly feel useless at times.". Based on the sum of the scores – accounting for subtractions in the case of reverse-coded questions – I could measure the self-esteem positioning of each individual on the Rosenberg Self-Esteem Scale. Finally, once the ID used to make the payment had been created, participants filled out a demographic questionnaire, providing their age, gender identity, academic cycle and year, and current GPA.

In the treatment groups, the procedure was similar, except for one moment: the disclosure of round predictions. Initially, subjects under conditions T3 and T4 received information on the performance of a past and similar cohort from the comparing control group. This included not only the average number of correctly counted tables but also a graph showing the distribution of that group's performance in percentage terms. Figure 1 illustrates the sequence of events described.

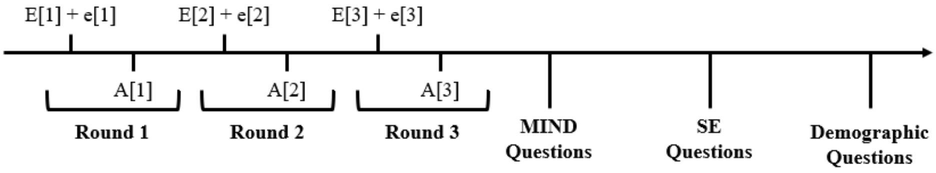


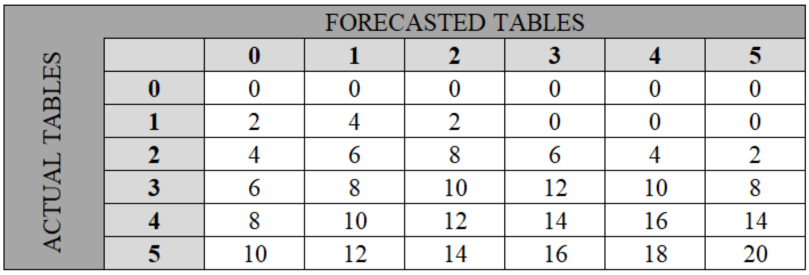
Figure 1 – Sequence of events during the experiment.

3.4. Payment Scheme

At the start of the experiment, participants were guaranteed a €5 show-up fee. For the variable payment, a dual incentive structure was designed based on Caplan et al. (2017) to align participants' motivations with real performance. Thus, in each round, the variable payment, in credits, depended not only on the number of tables in which the zeros were counted correctly, but also on the participants' *ex-ante* forecasts regarding their own performance.

To avoid hedging arguments as highlighted by Charness et al. (2016), the payoff was dependent only on a subset of choices. At the time of payment, each participant rolled a virtual 3-sided dice, corresponding to each of the three rounds, to determine which one would be paid. The credits collected were then converted into Euros, paid in Gift Cards. From

Figure 2, it is visible that depending on the accuracy of each participant, variable earnings could range from €0 to €20, with a total final value between €5 and €25, including the show-up fee. For example, in the table below, the value 20 corresponds to someone who forecasted counting 5 tables correctly, and actually counted them right, revealing both the best performance possible and perfect accuracy. It is important to note that whenever rounding was necessary due to a lack of cards, participants always benefited.



		FORECASTED TABLES					
		0	**1**	**2**	**3**	**4**	**5**
ACTUAL TABLES	**0**	0	0	0	0	0	0
	1	2	4	2	0	0	0
	2	4	6	8	6	4	2
	3	6	8	10	12	10	8
	4	8	10	12	14	16	14
	5	10	12	14	16	18	20

Figure 2 – Variable payment scheme table.

3.5. Procedures

The study was conducted at the Behavioural Research Lab[4] (XLAB) located at the Lisbon School of Economics and Management (ISEG). Participants were recruited through the Online Recruitment System for Economic Experiments (ORSEE), existing XLAB partnerships and initiatives, and posters displayed around ISEG. The subject pool was composed of people studying economics, management, finance or related fields.

To be eligible to participate, individuals had to meet three inclusion criteria: (a) be a university student; (b) understand English; and (c) not have previously taken part in the pre-test. Students self-selected the session that best suited their schedule by signing up for one of the available timeslots. The day before their session, I, the lab manager at the time, sent a reminder via email. The sessions were held over six days in May 2024.

3.6. Ethical Considerations

The approval for this study was obtained by the Ethics Committee from ISEG on May 8, 2024 (Research Ethics Approval No. 06/2024[5]). Upon arrival, participants were presented with an informed consent form on their monitor screens, giving them the option to either begin the study or withdraw without any consequences, aside from not receiving the participation fee.

During the study, participants were supervised by me and, after completing it, they were directed to the assistance room for payment. The study was implemented using the Qualtrics XM platform, which has limitations in dynamically updating and displaying cumulative earnings. Consequently, to ensure participants' anonymity a colleague

[4] A photo of the laboratory is shown in Appendix D.

[5] For a scanned version, see Appendix E.

conducted the process using an Excel sheet, calculating the amount based on an ID provided by each participant. The payments were made immediately after participation using Gift Cards.

3.7. Hypotheses and Analytical Framework

When designing the experiment, I intended to test specific hypotheses to test the impact of my intervention. By collecting the experimental data, I was able to analyse each participant's performance, expectations, and individual characteristics according to the condition to which they were assigned. Thus, it was possible to test the following hypotheses regarding the impact of information on the performance of a previous cohort:

Hypothesis 1: *The intervention has immediate (and positive) effects in reducing forecasting errors.*

Charalambous & Charalambous (2023) showed that the effects of some interventions, in treatment groups, can be seen immediately after the intervention. To do this, a cluster-randomised trial design was used. For those in the treatment group, 80-minute lessons were administered to help students develop their ability to formulate and use mathematical models to solve real-world problems. The study measured student performance before, immediately after and two months after the intervention. The immediate effects observed included significantly higher problem-solving performance in the treatment compared to the control group. Specifically, fifth graders in the experimental group performed better than sixth graders in the control group.

Hypothesis 2: *Participants in the treatment groups are more accurate when forecasting their own performance than those in the control groups in all rounds.*

Hypothesis 2.a): *Participants in the treatment groups are, on average, more accurate when forecasting their own performance than those in the control groups.*

Sabater-Grande et al. (2022) provided evidence that students generally overestimate their performance both when they set grade targets before exams and when they make post dictions immediately after exams. In the treatment groups, the incentives helped them make more accurate predictions about their own performance in all rounds, regardless of the students' inherent cognitive abilities.

Hypothesis 3: *Participants in the treatment groups are more accurate when forecasting their performance in relation to their peers than those in the control groups in all rounds.*

Tirso et al. (2019) found that top performers were not very confident about their relative position, despite accurately predicting their own grades. In turn, bottom-performing students were consistently overconfident about their grades and percentile ranks. Even after receiving feedback on the actual performance of the class, the underconfidence of the high-performing students in their relative position remained.

Hypothesis 4: *Participants in the treatment groups reveal greater improvement in the accuracy of their predictions about their performance in the final round compared to the first round than those in the control groups.*

Hypothesis 4.a): *Participants in the treatment group show greater improvement in the accuracy of their predictions from Round 1 to Round 2 and from Round 2 to Round 3 than those in the control groups.*

Magnus & Peresetsky (2018) reinforced the idea that students generally display overconfidence in their grade expectations. At the same time, students adjusted their expectations based on feedback and past performance. In this way, it was shown that participants in the treatment group showed a more consistent and significant improvement in their prediction accuracy from one round to the next, compared to participants in the control groups, who did not benefit from the feedback and learning opportunities provided in the treatment conditions.

Hypothesis 5: *Participants in the treatment groups with an implemental mindset give more accurate forecasts about their performance compared to participants in the control groups.*

Hügelschäfer & Achtziger (2014) analysed the impact of deliberative and implemental mindsets on confidence levels in judgment and decision-making tasks. Moreover, the study observed that mindset affects the sensitivity to anchoring effects. For instance, the implemental mindset reduces the influence of extrinsic factors on judgment accuracy. In the end, it was found that adopting an implementation mindset in treatment groups can lead to more accurate self-evaluations and performance predictions than in control groups.

Hypothesis 6: *There is a stronger relationship between self-esteem and average absolute delta in the control groups compared to the treatment groups.*

Kolovelonis & Goudas (2018) revealed how self-perceptions, including aspects of self-esteem, influence the accuracy of forecasts. It is suggested that interventions designed to improve self-regulated learning, such as clearly defining goals and providing feedback on mismatches between expected and actual performance, can help students develop a more realistic self-assessment. However, in the absence of such interventions (as in the control groups), the relationship between self-esteem and prediction accuracy appears to be stronger, as students show greater confidence in their inherent self-perceptions.

4. RESULTS

4.1. Descriptive statistics

With a total of 92 participants completing the study, no observations were excluded from the analysis. Table III below presents the demographic characteristics of the sample. The majority of the students were between 20 and 21 years old, accounting for 54.34% of the total sample. Overall, the pool was relatively gender-balanced, with 57.61% of participants being female. However, this distribution is not as balanced across the different experimental conditions. Regarding the academic year, 42.39% of the students reported being in their third year of their bachelor's, making it the most reported year across all conditions except for T3. In terms of academic performance, the GPA disclosed by the students suggests that, up to this point in their university life, they have generally experienced a median level of performance.

TABLE III

DEMOGRAPHIC CHARACTERISTICS OF THE SAMPLE

	Treatments				Combined
	T1 (N=23)	T2 (N=23)	T3 (N=23)	T4 (N=23)	(N=92)
Panel A: Age					
18	4.35%	13.04%	0%	4.35%	5.43%
19	8.7%	0%	13.04%	0%	5.43%
20	30.43%	26.09%	8.7%	30.43%	23.91%
21	17.39%	21.74%	47.83%	34.78%	30.43%
22	8.7%	4.35%	8.7%	8.7%	7.61%
23	13.04%	8.7%	8.7%	8.7%	9.78%
24	4.35%	4.35%	0%	13.04%	5.43%
25	0%	0%	0%	0%	0%
26	8.7%	4.35%	4.35%	0%	4.35%
27	0%	4.35%	4.35%	0%	2.17%
28	0%	8.7%	0%	0%	2.17%
29	0%	0%	0%	0%	0%
>30	4.35%	4.35%	4.35%	0%	3.26%
Panel B: Gender Identification					
Female	39.13%	60.87%	65.22%	65.22%	57.61%
Male	60.87%	39.13%	34.78%	34.78%	42.39%
Non-binary	0%	0%	0%	0%	0%
Panel C: Current University Cycle and Year					
1st year Bachelor's	4.35%	13.04%	4.35%	8.7%	7.61%
2nd year Bachelor's	13.04%	4.35%	21.74%	17.39%	14.13%
3rd year Bachelor's	60.87%	43.48%	26.09%	39.13%	42.39%
4th year Bachelor's	3.26%	0%	0%	0%	13.04%
1st year Master's	0%	4.35%	30.43%	13.04%	11.96%
2nd year Master's	17.39%	26.09%	13.04%	4.35%	15.22%
1st year PhD	4.35%	0%	4.35%	0%	2.17%
2nd year PhD	0%	4.35%	0%	0%	1.09%
3rd year PhD	0%	4.35%	0%	4.35%	2.17%
Panel D: Current Grade Point Average (GPA)					
10	0%	0%	4.35%	0%	1.09%
11	0%	0%	0%	0%	0%
12	4.35%	21.74%	8.7%	8.7%	10.87%
13	13.04%	17.39%	17.39%	34.78%	20.65%
14	34.78%	8.7%	13.04%	21.74%	19.57%
15	34.78%	13.04%	17.39%	17.39%	20.65%
16	4.35%	21.74%	13.04%	4.35%	10.87%
17	4.35%	13.04%	17.39%	8.7%	10.87%
18	4.35%	4.35%	8.7%	4.35%	5.43%
19	0%	0%	0%	0%	0%
20	0%	0%	0%	0%	0%

Figure 3 displays the earnings if all rounds had been paid, in Euros, under each condition. T1 has a median earning of €33 with an interquartile range (IQR) of 15, indicating a relatively tight clustering around the median. T2, with a median earning of €41 and an IQR of 23, shows the highest variability in earnings. T3 has the highest median earning at €43, with a moderate IQR of 19, reflecting a broader spread compared to T1 but narrower than T2. T4's median earning is €41, similar to T2, but with a smaller IQR of 14. The minimum and maximum values show that T3 has the widest range of earnings, from €5 to €63 (and the maximum would be €65), while T1 has a narrower range from €15 to €55. The mean earnings for T1, T2, T3, and T4 are €33.3, €40.0 and €39.8, respectively, suggesting that the average earnings follow a similar trend to the medians.

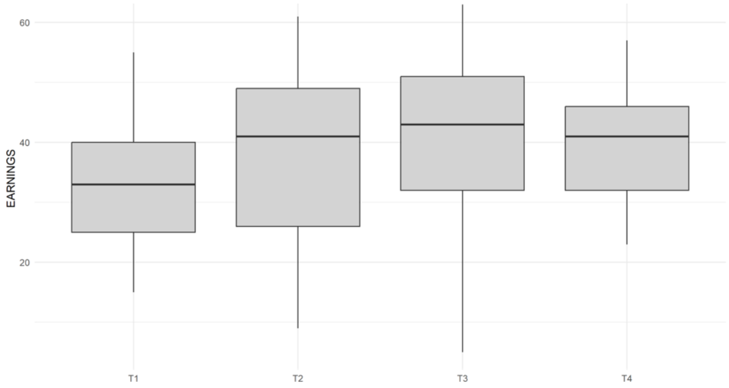


Figure 3 – Total earnings considering all rounds, per condition.

Table IV presents global descriptive statistics on the number of tables the students expected to count correctly on the three occasions they were asked, as well as the number of tables they ended up counting correctly. On average, students expected to count more tables correctly (3.46) than they actually did (3.18). The change in students' expectations was relatively small, with an average increase of 0.08 between the first and second tasks and 0.12 between the second and third tasks. Actual performance, however, showed, on average, a more significant improvement between the first and second tasks (0.72), but minimal change between the second and third tasks (0.02). This reveals that, on average, students expressed overconfidence, and although they improved their mean performance over time, this improvement was more pronounced at the beginning. Several graphs relating the variables of expected and actual performance can be found in Appendix F.

TABLE IV

GLOBAL DESCRIPTIVE STATISTICS ON EXPECTED AND ACTUAL PERFORMANCE

	Mean	Median	Mode	Min	Max	SE
Expected Performance						
E[1]	3.37	3	3	2	5	0.89
E[2]	3.45	3	3	2	5	0.82
E[3]	3.57	4	3	1	5	0.98
Average E	3.46	3.33	3	2	5	0.71
Actual Performance						
A[1]	2.70	3	3	0	5	1.16
A[2]	3.41	4	4	0	5	1.29
A[3]	3.43	4	5	0	5	1.39
Average A	3.18	3.33	3.33	0.33	5	0.99
Change in Expected Performance						
E[2] - E[1]	0.08	0	0	-2	2	0.96
E[3] - E[2]	0.12	0	0	-2	2	0.80
Change in Actual Performance						
A[2] - A[1]	0.72	1	1	-3	4	1.31
A[3] - A[2]	0.02	0	1	-3	4	1.51

In the next subsections, under/overconfidence is measured both in absolute (relative to one's own performance) and relative (positioning relative to peers) terms. The metrics used were as follows:

(2) Absolute Δ_n = E[n] - A[n],

(3) Relative Δ_n = e[n] - a[n], where e[n] was converted into quartiles as mentioned in subsection *3.2. Variables*.

(4) Average absolute/relative $\Delta = \frac{\Delta 1+\Delta 2+\Delta 3}{3}$

Thus, taking the absolute Δ as an example, a negative value would indicate underconfidence (i.e. underestimation of one's ability), while a positive value would indicate overconfidence (i.e. overestimation of one's ability).

4.2. Immediate effects of the intervention

As mentioned in the subsection *3.1. Treatments*, the first round was the same for all participants. Therefore, and only this time, it is possible to combine conditions T1 with T2 and conditions T3 with T4 to assess the effectiveness of the treatment[6]. Figure 4 shows

[6] Appendix A includes the tests that prove there are no statistically significant differences for round 1.

the boxplots that were created to visualise the distribution of absolute Δ_1 for the T1+T2 and T3+T4 groups. The main difference is in the median absolute Δ_1: For the control groups, the participants were overconfident about their own performance (with a median value equal to 1), while in the control group, the mean predictions were accurate.

To test Hypothesis 1, a Wilcoxon Rank-Sum test[7] with continuity correction was carried out. The results of the test were as follows: $W = 1420$ ($p = 0.004 < 0.01$). Thus, this result indicates a statistically significant difference in prediction errors between the control and treatment groups. Returning to the graph, we also confirm that this difference implies a smaller absolute Δ_1 in the treatment groups, so the hypothesis that treatment has positive immediate effects is not rejected.

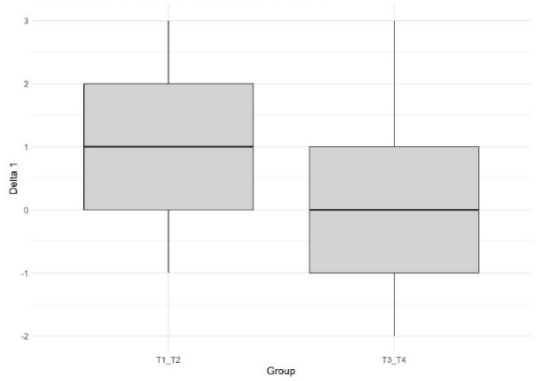


Figure 4 – Absolute Δ_1 for the control and treatment groups.

4.3. Absolute Under/Overconfidence

The boxplots in Figure 5 and Figure 6 depict the distribution of absolute Δ values per round for all conditions. At first glance, the most prominent result is related to T4, since the median absolute Δ was 0 in all rounds. But first, focusing on the comparisons between T1 and T3, in round 1, T1 had an IQR of 2.5 and a maximum value of 4, indicating a high variability in overconfidence. The median value of the absolute Δ_1 was 1.22. In contrast, T3 had a slightly narrower IQR of 2, with a median absolute Δ_1 of 0, showing less volatility as well as greater accuracy in the forecast. Furthermore, in round 2, the median absolute Δ_2 in T1 was equal to the third quartile value in T3 (with a value of 0), implying that the median Δ_2 in T1 was higher than most of the values in T3, and, consequently, that

[7] All statistical tests performed were of this type because of the nature of the data, the need to compare paired samples, the sample size and the lack of sensitivity to outliers.

T1 participants had higher overconfidence. In round 3, T1 had an IQR of 1, half the value presented in T3. In fact, this was the only round in which the IQR was lower for T1 than for T3. Participants in T1 showed a median absolute Δ_3 of 1 (contrasting with 0 for T3). Note that the point above "Abs. Δ_3" in the graph indicates an outlier, while the median matched the minimum, showing an asymmetrical distribution.

For T2 and T4, in the control group, round 1 had an IQR of 2 and a median absolute Δ_1 of 1, while in T4, the IQR was narrowed to 1, with a median of 0. In round 2, T2 and T4 had an IQR of 1 with a median absolute Δ_2 of 0, indicating similar volatility. In round 3, the median absolute Δ_3 in T2 was equal to the first quartile value in T4. When analysing the pattern of descending bars in T2 over the rounds, we observed a continuous reduction in the median absolute Δ values, so it can be deducted that participants in this condition went from being overconfident about their own performance to underconfident.

Hypothesis 2 claims, in other words, that participants in the treatment groups have a lower absolute Δ than participants in the control groups in all rounds. The results of the Wilcoxon Rank-Sum test provide evidence to reject it. For T1 vs T3, the significant differences in rounds 1 ($p = 0.017 < 0.05$) and 3 ($p = 0.000 < 0.001$) indicate that T3 participants have less overconfidence. However, no statistically significant difference was found in round 2 ($p = 0.167$). Furthermore, when comparing T2 to T4, no statistically significant differences were found (p-values of 0.096, 0.643 and 0.436 for rounds 1 to 3, respectively), suggesting similar levels of overconfidence. Regarding Hypothesis 2.a), the test results, a similar procedure showed a statistically significant difference in the average forecast accuracy between T1 and T3 ($p = 0.002 < 0.01$), which is supported by the lower median absolute Δs shown above for rounds 1 and 3. However, again, for T2 against T4 no statistically significant difference was found ($p = 0.259$). Thus, the claim that students in the treatment groups make, on average, more accurate forecasts about their performance is only partially rejected.

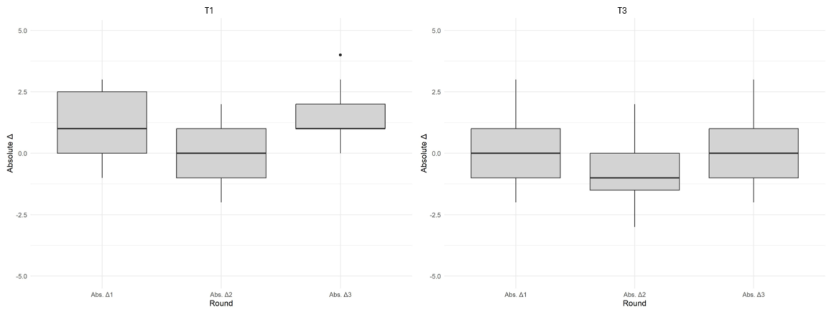


Figure 5 – Absolute Δ per round in T1 and T3.

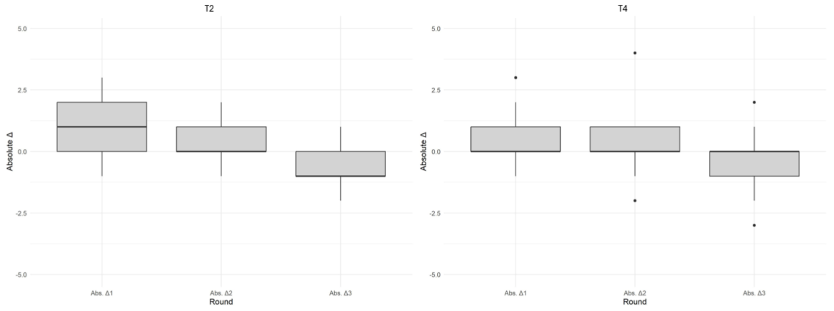


Figure 6 – Absolute Δ per round in T2 and T4.

Overall, Figure 7, offers results that go in line with past literature on top-performing students (Ehrlinger et al., 2008; Nederhand et al., 2020; Tirso et al., 2019). In this study, students under group "A" either suffered from underconfidence or were able to accurately predict their performance. In T1 they showed overconfidence with an average absolute Δ of 0.333, while participants in T3 showed a lack of confidence, as this value was negative at -0.875. T2 and T4 both revealed average absolute Δs below 0 (-0,167 in T2 vs -0,278 in T4).

In the comparison between T1 and T3, bottom performers showed overconfidence in both conditions, with T1 having an average absolute Δ of 1.29 and T3 having a higher average Δ of 1.58. One of the reasons may be due to their lack of capacity, which makes it difficult for them to learn, and constrains them from adjusting their confidence levels as they receive new information (Karaca et al., 2023). In the graph comparing T2 with T4, students classified as "D" in both groups had overconfidence levels close to 1.

For the middle groups, the intervention inevitably made the average absolute Δ less positive, which may or not be a good indicator. Looking at group "B" at T4, who were already underconfident, this intervention increased even more the size of the error. It is therefore necessary to consider how the various groups are impacted before making a decision about whether to implement measures aimed at reducing overconfidence. What for some may be a help in the right direction, for others may further exacerbate their difficulties in forming real estimates.

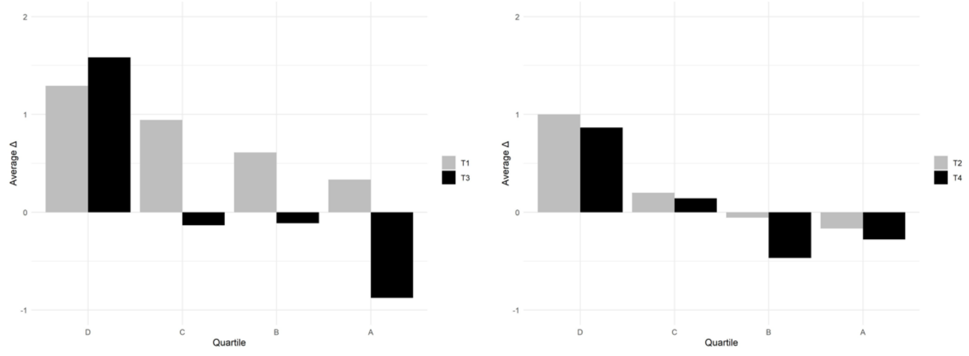


Figure 7 – On the left, the average absolute Δ (for all rounds) by grade, for T1 and T3. On the right, the average absolute Δ (for all rounds) by grade, for T2 and T4.

4.4. Relative Under/overconfidence

At first glance, looking at Figure 8 it seems that the intervention is marked by a relative underconfidence that fades with practice. However, one should be cautious with the validity of the conclusions taken as these forecasts were not considered for variable earnings. In T1, the medians for relative Δ_2 and relative Δ_3 were 0, indicating that individuals generally had a realistic perception of their performance relative to their peers. The variability in T1 was considerable, with wide IQRs that ranged from 2 in Round 1 to 2.5 in the other rounds. In contrast, T3 exhibited more pronounced underconfidence, particularly in relative Δ_1 and relative Δ_2, with medians of -1 each and IQRs of 2 each. This attested that, individuals in T3, tended to underestimate their performance relative to their peers in these rounds. For relative Δ_3, the median was 0 with an IQR of 1.5, suggesting that participants became more overconfident (as occurred in T1) and were capable of adjusting their perceptions closer to reality over time.

Regarding Figure 9, T2 showed a balanced perception of performance with medians around 0 for relative Δ_1 and relative Δ_2, but a slight underconfidence in round 3 (with a median of -1). The IQRs indicated some variability, with the widest range in round 1 (2.5) and narrowing in subsequent rounds. T4, on the other hand, showed a more consistent pattern with medians of relative Δs at 0 for all rounds, indicating a globally realistic or slightly conservative self-assessment. The IQRs were tinner than in T2, particularly in relative Δ_2 and relative Δ_3 (1 each), which contributed to better accuracy in predicting the percentile positioning.

To examine the impact of the intervention on the relative predictions I performed the Wilcoxon Rank-Sum test. Comparing T1 with T3, there were only significant differences in round 2 ($p = 0.022 < 0.05$) and round 3 ($p = 0.003 < 0.01$), implying that participants in T1 were able to better evaluate their performance in relation to that of their peers while in T3 they tended to be underconfident. For the first round, no statistically significant difference was encountered ($p = 0.453$). For T2 and T4 the results showed no statistically significant differences (p-values of 0.698, 0.876, and 0.649 for rounds 1 to 3, respectively). Thus, Hypothesis 3, is rejected and, to some extent even reversed, since the only statistically significant results were for two rounds in which the participants in the control group, T1, were the most accurate.

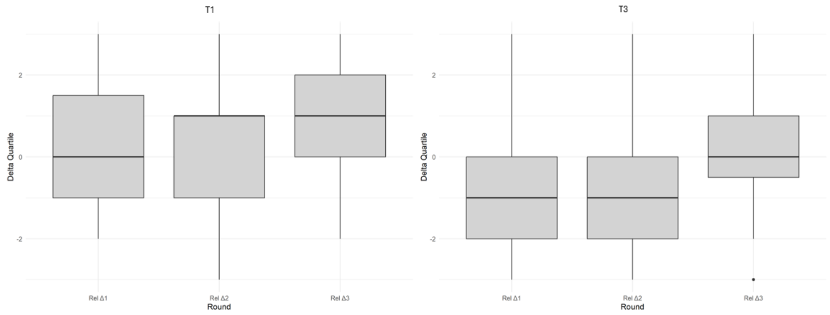


Figure 8 – Relative Δ per round in T1 and T3.

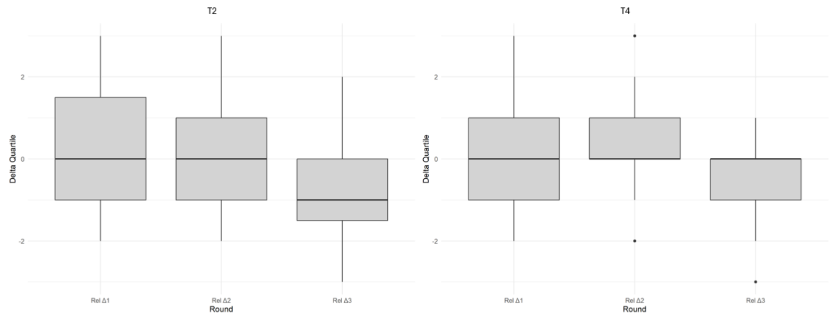


Figure 9 – Relative Δ per round in T2 and T4.

4.5. Learning

Figure 10 provides evidence of more consistent and stable learning for the treatment groups. Between the first two rounds, for T1 there was an average absolute Δ difference of -0.739, indicating a substantial decrease in error. In contrast, T3 showed a much smaller reduction with an average absolute Δ difference of -0.044. Thus, T1 had the most substantial improvement in error reduction compared to T3, although both groups became more accurate. In contrast, between rounds 2 and 3, T1 had an average absolute Δ difference of 0.783, signalling an increase in error. However, T3 continued to present a decrease in error with an average absolute Δ difference of -0.217. So, T1 participants showed a regression in learning, while T3 participants demonstrated an improvement, albeit a smaller one. Regarding overall learning over the course of the study, T1 had an average absolute Δ difference of 0.044, so in total, there was little loss of learning in the ability to make accurate predictions. On the other hand, as expected, T3 showed a decrease in error with an average absolute Δ difference of -0.261.

Comparing T2 with T4, between rounds 1 and 2 there was an average absolute Δ difference of -0.348 among T2 participants, i.e. the error decreased. T4, however, showed an increase in error with an average absolute Δ difference of 0.130. In the next comparison, between rounds 2 and 3, the roles were reversed. In T2 there was an average absolute Δ difference of 0.304 while in T4 this value was -0.0435. Thus, T4 participants' were the ones showing better predictive skills. Finally, between the first and last rounds, students in the T2 condition showed an average absolute Δ difference of -0.0435, suggesting that they had returned to improve their accuracy. T4, however, showed an increase in error with an average absolute Δ difference of 0.087, so, even slightly, they raised their errors.

To check hypotheses 4 and 4.a), I run a Wilcoxon test to compare the differences in the evolution of forecast errors. When comparing the results from round 1 to round 3, there was no statistically significant difference in the accuracy of T1 participants compared to T3 participants over the entire course of the rounds (p = 0.334). For T2 vs T4, the p-value was 0.883, suggesting a similar conclusion. Therefore, hypothesis 3 is rejected, since in none of the conditions did the forecasts become statistically significant more accurate. Regarding Hypothesis 3.a), only from round 2 to 3 and comparing T1 with T3 were found statistically significant differences ($p = 0.002 < 0.01$), with T3 participants showing a decrease in error that was not verified in T1. For the same rounds, for T2 vs T4 the p-value was 0.883. From round 1 to round 2, the p-values were also not statistically significant (T1 vs T3: p = 0.078, T2 vs T4: p = 0.143), meaning that this hypothesis is only partially not rejected.

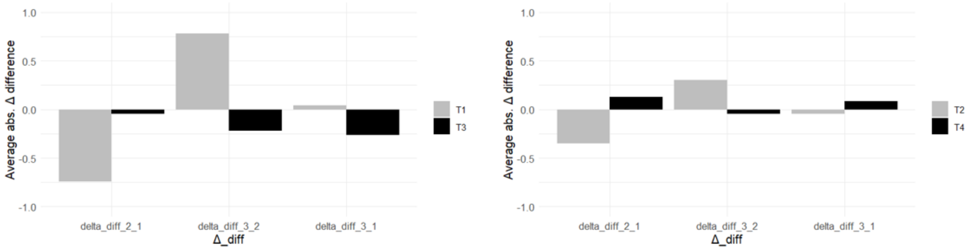


Figure 10 – On the left, the average absolute Δ difference between rounds, for T1 and T3. On the right, the average absolute Δ difference between rounds, for T2 and T4.

Expected performance decreased across rounds, with E[1] dropping by 1.00, E[2] by 0.08, and E[3] by 0.35 from T1 to T3. Not surprisingly, the average expected performance fell by 0.48 as well. Conversely, the actual performance improved: A[1] increased by 0.08, A[2] by 0.39, and A[3] by 1.05. The average actual performance rose by 0.51. From T2 to T4, the expected performance values showed minimal change with E[1] decreasing by 0.31 and E[2] by 0.22, while E[3] increased by 0.35. This resulted in an average expected performance that decreased slightly by 0.06. Once again, the actual performance values showed a slight improvement: A[1] increased by 0.26, A[2] decreased by 0.05, and A[3] increased by 0.17. The average actual performance increased by 0.14. In summary, Table V demonstrates that the reduction in the gap between expected and actual performance was largely due to an improvement in actual performance, as it was always larger than the reduction in the average expected performance.

TABLE V

MEAN VALUES OF EXPECTED AND ACTUAL PERFORMANCE ACROSS CONDITIONS

	T1	T2	T3	T4
Expected Performance				
E[1]	3.91	3.48	2.91	3.17
E[2]	3.65	3.39	3.57	3.17
E[3]	3.83	3.30	3.48	3.65
Average E	3.80	3.39	3.32	3.33
Actual Performance				
A[1]	2.70	2.52	2.78	2.78
A[2]	3.70	2.96	4.09	2.91
A[3]	2.30	3.96	3.35	4.13
Average A	2.90	3.14	3.41	3.28

4.6. Mindset

The average absolute Δ values for each mindset are illustrated in Figure 11 to understand the differences in forecasting accuracy. For the implemental mindset, the average absolute Δ was higher for T1 (0.910) compared to T3 (-0.379), indicating that the presence of information shifted an overprediction to underprediction. For the deliberative mindset, the average absolute Δ was also higher for T1 (0.933) than for T3 (0.306). In fact, in terms of absolute values, predictions were more accurate in T3, regardless of the mindset. In the comparison between T2 and T4, for the implemental mindset, the average absolute Δ was slightly higher for T4 (0.278) than for T2 (0.214), indicating a decrease, even minor, in the forecast accuracy for T4. In contrast, for the deliberative mindset, the average absolute Δ was higher for T2 (0.407) compared to T4 (0.167).

To statistically assess these differences, a Wilcoxon Rank-Sum test was performed. Results showed a statistically significant difference in the forecast accuracy between T1 and T3 for the implemental mindset ($p = 0.001 < 0.01$). However, there was no statistically significant difference between T1 and T3 for the deliberative mindset ($p = 0.246$), indicating the treatment had no significant effect. No statistically significant differences were found between T2 and T4 for either mindset (p-values equal to 0.898 and 0.176, respectively), which may lead us to suspect that observed differences may be due to random chance rather than a true treatment effect.

These results only partially reject Hypothesis 5, indicating that for participants with an implemental mindset, the average absolute Δ is significantly lower, and more accurate, in T3 compared to T1, but not in other comparisons. The intervention, that consisted on providing data about the performance of a previous cohort, is key to understanding these findings. Participants with an implemental mindset may have benefited from having a clear and graphical representation of an information relevant in this context, as they are commonly characterised as being closed-minded. Ultimately, this idea that they could have seen the information provided as anchors may be further investigated.

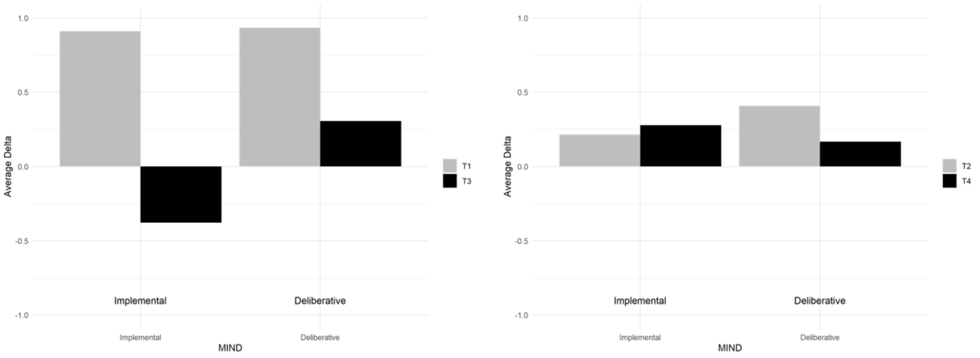


Figure 11 – On the left, the average absolute Δ depending on the mindset, for T1 and T3. On the left, the average absolute Δ depending on the mindset, for T2 and T4.

4.7. Self-esteem

To analyse the relationship between SE and average absolute Δ in the control and treatment groups, a combination of Pearson correlation analysis and linear regression was employed. Figure 12 presents the scatter plots generated to inspect the relationship between SE and average absolute Δ for each group, with a linear regression line added to each to show the trend. For each condition, the Pearson correlation coefficient and its associated p-value were calculated. This coefficient, which ranges from -1 to 1, indicates the degree of linear relationship between the variables, with values closer to 1 (-1) implying a strong positive (negative) correlation, and values around 0 suggesting no correlation.

In T1, there was a weak negative correlation (-0.092) between SE and average absolute Δ, which was not statistically significant (p = 0.676). In other words, a higher average absolute Δ was associated with lower SE, although the relationship was not meaningful. Similarly, T3 participants exhibited a weak negative correlation (-0.102) between SE and average absolute Δ, also not statistically significant (p = 0.644), indicating again no meaningful relationship between SE and average absolute Δ in this condition. In contrast, T2 showed a moderate positive correlation (0.39) between SE and average absolute Δ, which was marginally statistically significant (p = 0.066). Finally, in T4 there was a weak positive correlation (0.174) which was not statistically significant (p = 0.428), indicating no meaningful relationship in this group either.

After, for each condition, a simple linear regression model was fitted using SE as the response variable and average absolute Δ as the predictor. I opted for this tool to quantify the relationships with a specific equation, have detailed information on the statistical significance of both the slope and the intercept, and understand the proportion of variance explained by the model. In T1, the model explained 0.8% of the variance in SE, with a non-statistically significant positive relationship between SE and average absolute Δ (coefficient = -0.813, p = 0.676). Not surprisingly, T2's model explained only 1.0% of the variance in SE, with a non-statistically significant relationship between SE and average absolute Δ (coefficient = -0.921, p = 0.644). In the T3 group, the model explained 15.2% of the variance in SE, revealing a marginally statistically significant positive relationship between SE and average absolute Δ (coefficient = 2.207, p = 0.066). For T4, the model explained only 3.0% of the variance in SE, with a non-statistically significant relationship between SE and average absolute Δ (coefficient = 1.343, p = 0.0.428).

After looking at the multiple linear regression with an interaction term to test if the relationship between SE and average absolute Δ differs between control (T1 and T2) and treatment (T3 and T4) groups, the findings do not support Hypothesis 6. So, we have evidence to reject it.The interaction between treatment and average absolute Δ was significant (coefficient = 3.234, p = 0.040 < 0.05), indicating that the relationship between SE and average absolute Δ differs between the control and treatment groups. More specifically, in the control groups, there was a negative relationship between SE and average absolute Δ (slope = -1.170), whereas in the treatment groups, this relationship was positive (slope = 2.070). If we check the absolute values, the strength of the relationship was greater in the treatment than in the control groups.

A possible explanation for this may be that the introduction of new data that could have been viewed as a tool to make comparisons between the self and others did not lead to a reduction in the influence of self-esteem, but changed the way self-esteem influenced performance estimates. Alternatively, the treatment condition may have induced a form of self-affirmation or competitiveness, in which participants with higher self-esteem responded more positively to the information provided.

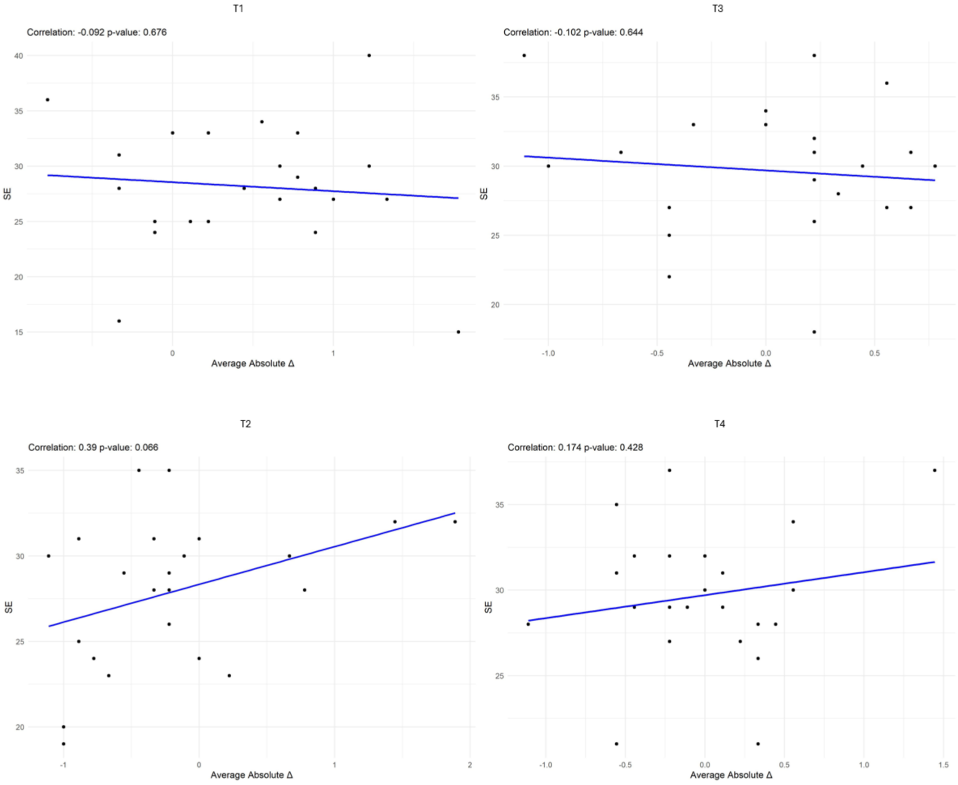


Figure 12 – The correlation between self-esteem (SE) and average absolute Δ in each of the conditions.

4.8. Regression Models

When doing these regressions, the response variable was "Average absolute Δ", henceforth known as avg_delta, as it is the measure of under/overconfidence and my primary goal in this research is to reduce overconfidence. For that, in my initial approach to understanding the factors influencing the avg_delta, I employed a multiple linear regression model without interaction terms[8]. COND was not treated as a categorical variable with multiple levels but rather as a continuous or ordinal variable. The regression is explicitly written in this section as:

[8] All this section was replicated with regressions considering the interaction term "CYL x YEAR", as these variables are related, in Appendix G. However, as the significant predictors were the same and the best model performed no better than model 4, I stuck with the models without interaction terms.

(5) $avg\_delta_i = \beta_0 + \beta_1 COND_i + \beta_2 MIND_i + \beta_3 SE_i + \beta_4 GEN_i + \beta_5 AGE_i + \beta_6 CYL_i + \beta_7 YEAR_i + \beta_8 GPA_i + \epsilon_i$

Through Table VI, it is noticeable that only GPA emerged as a statistically significant predictor of avg_delta at a 1% level (p = 0.008) while MIND (p = 0.064), SE (p = 0.087) and CYL (p = 0.066) showed marginal significance. The β_1 = -1.338 indicates that for each unit increase in COND, the avg_delta decreases by approximately 1.338 units, *ceteris paribus*.

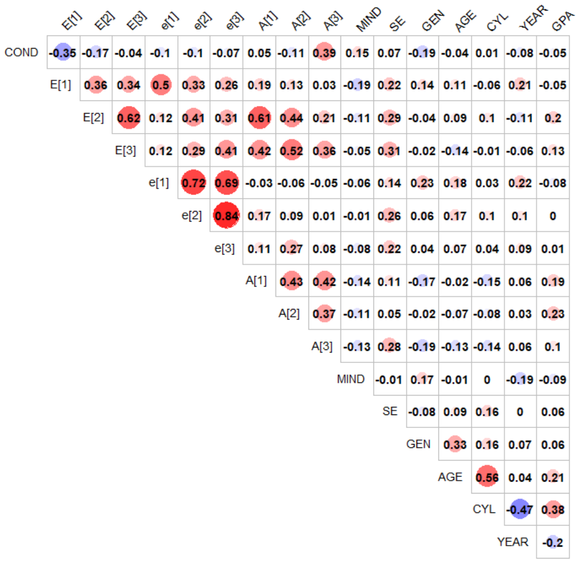
The model exhibited a Multiple R^2 value of 0.186, indicating that approximately 18.6% of the variability in avg_delta was explained by the predictors. The Adjusted R^2 value was 0.107, suggesting limited generalizability. For these reasons, it appeared that adding more terms that could capture the influence of performance could possibly lead to a more robust and accurate model.

Model 2 was a more complex model, as it included all the variables I could take from the experiment. The regression was expressed as:

(6) $avg\_delta_i = \beta_0 + \beta_1 COND_i + \beta_2 E[1]_i + \beta_3 E[2]_i + \beta_4 E[3]_i + \beta_5 e[1]_i + \beta_6 e[2]_i + \beta_7 e[3]_i + \beta_8 A[1]_i + \beta_9 A[2]_i + \beta_{10} A[3]_i + \beta_{11} MIND_i + \beta_{12} SE_i + \beta_{13} GEN_i + \beta_{14} AGE_i + \beta_{15} CYL_i + \beta_{15} YEAR_i + \beta_{17} GPA_i + \epsilon_i.$

The same table revealed improvements both in the explanatory power and model fit. Several predictors, including A[1], A[2] and A[3] were highly significant (p < 0.001). The Multiple R^2 value of 0.866 indicated that the enhanced model explained 86.6% of the variability in avg_delta a substantial improvement over the initial model. The Adjusted R^2 value was 0.835, suggesting that the model generalizes well to new data, although there may be some variables that make the model worse than expected.

Nevertheless, a correlation matrix was generated to assess multicollinearity among the predictors. Figure 13 is a visual representation of the Pearson correlation coefficients between the various predictors included in regression model 2. It should be read as follows: positive (negative) correlation coefficients are represented by red (blue) circles, where the size and intensity of the colour indicate the strength of the correlation. The values within the cells are the numerical correlation coefficients and are within the continuous interval [-1;1], with 1 meaning perfect positive correlation.



	E[1]	E[2]	E[3]	e[1]	e[2]	e[3]	A[1]	A[2]	A[3]	MIND	SE	GEN	AGE	CYL	YEAR	GPA
COND	-0.35	-0.17	-0.04	-0.1	-0.1	-0.07	0.05	-0.11	0.39	0.15	0.07	-0.19	-0.04	0.01	-0.08	-0.05
E[1]		0.36	0.34	0.5	0.33	0.26	0.19	0.13	0.03	-0.19	0.22	0.14	0.11	-0.06	0.21	-0.05
E[2]			0.62	0.12	0.41	0.31	0.61	0.44	0.21	-0.11	0.29	-0.04	0.09	0.1	-0.11	0.2
E[3]				0.12	0.29	0.41	0.42	0.52	0.36	-0.05	0.31	-0.02	-0.14	-0.01	-0.06	0.13
e[1]					0.72	0.69	-0.03	-0.06	-0.05	-0.06	0.14	0.23	0.18	0.03	0.22	-0.08
e[2]						0.84	0.17	0.09	0.01	-0.01	0.26	0.06	0.17	0.1	0.1	0
e[3]							0.11	0.27	0.08	-0.08	0.22	0.04	0.07	0.04	0.09	0.01
A[1]								0.43	0.42	-0.14	0.11	-0.17	-0.02	-0.15	0.06	0.19
A[2]									0.37	-0.11	0.05	-0.02	-0.07	-0.08	0.03	0.23
A[3]										-0.13	0.28	-0.19	-0.13	-0.14	0.06	0.1
MIND											-0.01	0.17	-0.01	0	-0.19	-0.09
SE												-0.08	0.09	0.16	0	0.06
GEN													0.33	0.16	0.07	0.06
AGE														0.56	0.04	0.21
CYL															-0.47	0.38
YEAR																-0.2

Figure 13 – Correlation matrix of variables represented in the regression model 2, using a color-coded and size-scaled bubble plot.

The matrix revealed a high correlation between e[2] and e[3] (0.84 > 0.8[9]), indicating multicollinearity issues. This level of multicollinearity can inflate the variance of coefficient estimates and destabilize the model. To address this problem, I decided to test two additional models to determine which variable to exclude to mitigate the multicollinearity problem:

Model 3

$$(7)\ avg\_delta_i = \beta_0 + \beta_1 COND_i + \beta_2 E[1]_i + \beta_3 E[2]_i + \beta_4 E[3]_i + \beta_5 e[1]_i + \beta_6 e[2]_i + \beta_7 A[1]_i + \beta_8 A[2]_i + \beta_9 A[3]_i + \beta_{10} MIND_i + \beta_{11} SE_i + \beta_{12} GEN_i + \beta_{13} AGE_i + \beta_{14} CYL_i + \beta_{15} YEAR_i + \beta_{16} GPA_i + \epsilon_i.$$

Model 4

$$(8)\ avg\_delta_i = \beta_0 + \beta_1 COND_i + \beta_2 E[1]_i + \beta_3 E[2]_i + \beta_4 E[3]_i + \beta_5 e[1]_i + \beta_6 e[3]_i + \beta_7 A[1]_i + \beta_8 A[2]_i + \beta_9 A[3]_i + \beta_{10} MIND_i + \beta_{11} SE_i + \beta_{12} GEN_i + \beta_{13} AGE_i + \beta_{14} CYL_i + \beta_{15} YEAR_i + \beta_{16} GPA_i + \epsilon_i.$$

[9] Values above 0.8 or below -0.8 were considered worrying.

After comparing the two models, model 4 manifested slightly better performance metrics, explaining 86.6% (marginally better than the 86.4% from model 3) of the variability in avg_delta. Even more, the root mean square error was also lower, with the value 3.701 against 3.723 from model 3, and the F-statistics was higher at 30.23 (against 29.81 in model 3). Finally, as both models continued to consider A[1], A[2] and A[3] as highly significant predictors, model 4 was the preferred choice.

TABLE VI

PREDICTORS OF AVG_DELTA

	Model 1		Model 2		Model 3		Model 4	
Predictors	β_i	p-Value	β_i	p-Value	β_i	p-Value	β_i	p-Value
(Intercept)	5.725	0.674	3.081	0.614	3.191	0.601	3.029	0.617
COND	-1.338	0.115	-0.068	0.880	-0.085	0.850	-0.070	0.875
E[1]	-	-	0.652	0.326	0.481	0.451	0.649	0.325
E[2]	-	-	0.920	0.285	0.743	0.376	0.982	0.214
E[3]	-	-	-0.651	0.317	-0.423	0.484	-0.675	0.288
e[1]	-	-	-0.052	0.257	-0.036	0.398	-0.050	0.257
e[2]	-	-	0.010	0.850	0.046	0.259	-	-
e[3]	-	-	0.046	0.334	-	-	0.052	0.140
A[1]	-	-	-3.662	2.95e-10***	-3.711	1.53e-10***	-3.654	2.15e-10***
A[2]	-	-	-3.169	1.82e-10***	-3.014	7.65e-11***	-3.189	4.15e-11***
A[3]	-	-	-2.189	3.00e-07***	-2.206	2.35e-07***	-2.190	2.44e-07***
MIND	3.637	0.064	1.259	0.155	1.116	0.200	1.290	0.135
SE	-0.341	0.087	-0.043	0.660	-0.040	0.684	-0.041	0.668
GEN	0.816	0.670	-0.131	0.887	-0.201	0.827	-0.143	0.876
AGE	-0.244	0.673	-0.130	0.621	-0.126	0.631	-0.159	0.621
CYL	4.527	0.067	-0.595	0.591	-0.639	0.563	-0.582	0.596
YEAR	0.990	0.479	0.570	0.361	0.529	0.395	0.580	0.347
GPA	-1.524	0.008**	0.116	0.656	0.106	0.682	0.115	0.655
Other Diagnostics								
Root MSE	8.665		3.725		3.723		3.701	
F-statistics	2.362	0.024	28.09	<2.2e-16	29.81	<2.2e-16	30.23	<2.2e-16
Multiple R^2	0.186		0.866		0.864		0.866	
Adj.R^2	0.107		0.835		0.835		0.837	

*p-Value < 0.05; **p-Value < 0.01; *** p-Value < 0.001

After selecting model 4, the next step involves evaluating the model's assumptions and diagnostics using Figure 14, which on the y-axis plots the residuals (the differences between observed and predicted values) against the fitted values (the predicted values) on the x-axis. When analysing it, four conclusions can be drawn. Firstly, the residuals display a non-random pattern around the horizontal blue line (residual = 0), forming an inverted U-shape. This indicates the presence of non-linearity in the data, creating doubt that the linear model may not fully capture the underlying relationship between the predictors and avg_delta.

Moreover, the residuals exhibit increasing variance with higher fitted values, revealing heteroscedasticity. In other words, the variability of the residuals is not constant across all levels of fitted values, violating one of the key assumptions of linear regression, which can appear through inefficient estimates and underestimated standard errors. Another issue concerns the several data points with large residuals, both positive and negative. These points may be potential outliers or influential observations that can disproportionately affect the model's parameter estimates. Not less importantly, the observed distribution shows some asymmetry, suggesting possible problems with the normality of residuals, which ideally should be symmetrically distributed around zero.

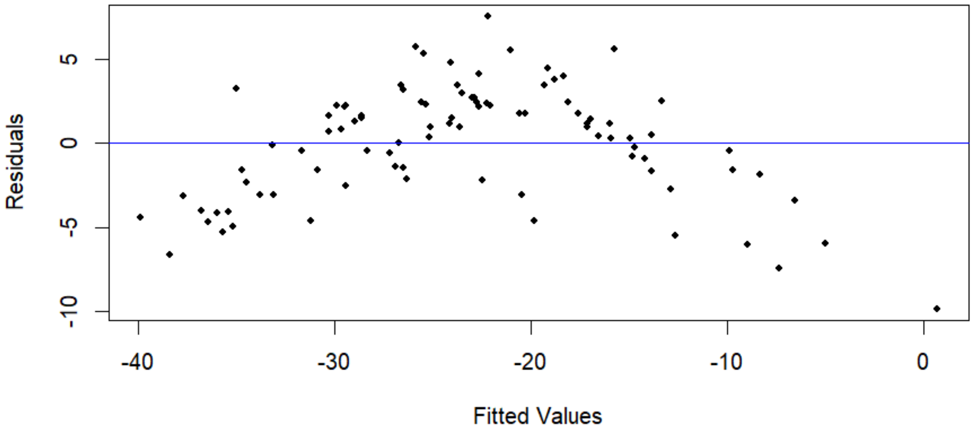


Figure 14 – Residuals vs Fitted Plot[10] from the regression model 4.

Starting from the last observation related to the normality of the model's residuals, the points in Figure 15 below represent the quantiles of the residuals against the

[10] Residuals and Fitted values use the same units as the dependent variable, avg_delta. In this case there are such high values due to the inclusion of variables related to the expected percentile in relation to peers, which could vary between 1 and 99.

theoretical quantiles of a standard normal distribution. In the central region of the graph, most of the points lie close to the 45-degree blue reference line and show being reasonably symmetric, which supports the assumption of normality. Despite this, at the extremes of the distribution (both the lower left and upper right), the points deviate significantly from the reference line. The presence of heavy tails implies that there are more extreme values in the residuals than would be expected under normality, which once again recommends the investigation of outliers.

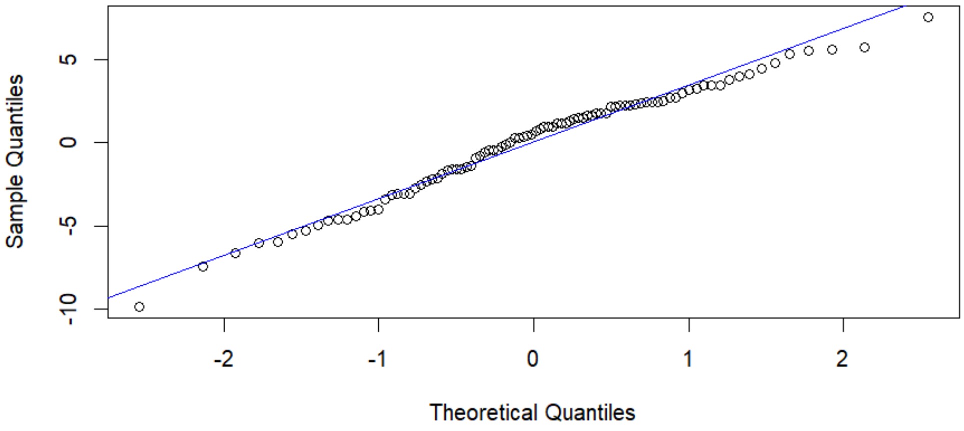


Figure 15 – Normal Q-Q Plot from the regression model 4.

Following on from the previous discussion, Figure 16 detects outliers and points with high leverage, those whose observation's predictor values are from those of other observations. Here, leverage values range from 0 to about 0.45[11]. The standardized residuals in this plot range from approximately -3 to 2. Points with large standardized residuals, particularly those above 2 or below -2, are potential outliers and should be treated with caution. Points highlighted with numbers (40, 64 and 74) have higher leverage and may be influential based on Cook's distance (the red contours), as a result, further investigation is advised.

[11] Observations with leverage close to 0.5 or higher are considered to have high leverage.

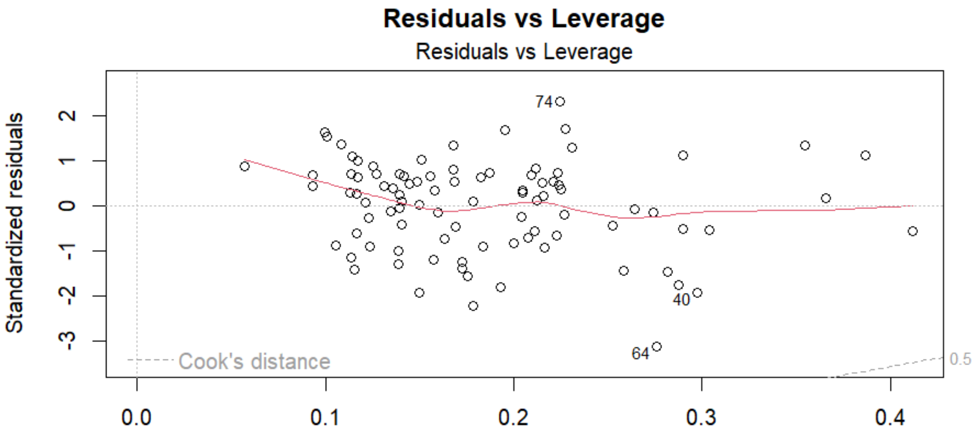


Figure 16 – Standardized residuals against the leverage of each observation in the regression model 4.

Turning to the relative importance of variables in the regression model, as depicted below, it is possible to visualize the contribution of each predictor to the model. The blue dashed line serves as a visual benchmark, helping to distinguish between variables that are above average in terms of their importance and those that are below average. As such, A[1], A[2], MIND and A[3] are likely the primary drivers of avg_delta, indicating that being in a deliberative or implemental mindset has a strong influence on the response variable. On the other hand, SE, e[1] and e[3] are negligible predictors and removing them may simplify the model. It is important to note that while, all the values depicted are positive, this does not imply that their relative importance in the model has a positive impact. The values merely indicate the magnitude of the importance of each variable, not the direction of their effect, which should be seen by looking at the coefficients of these variables in model 4 (in Table VI).

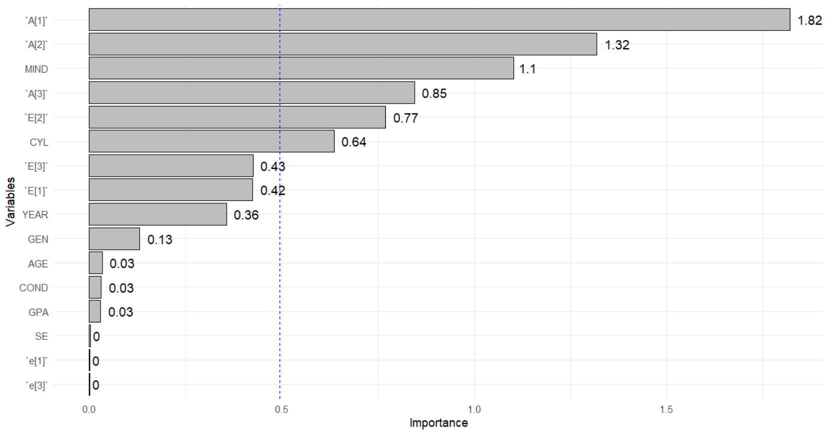


Figure 17 – Relative importance of the variables in the regression model 4.

In sum, the analysis of the variables influencing avg_delta reveals that actual performance in counting tables is crucial. A[1], A[2], and A[3] exhibit the following coefficients -3.654, -3.189, and -2.190, respectively, and their p-values are all significant at a 0.1% level. This indicates that higher actual performance in these rounds strongly predicts less overconfidence, or even underconfidence. Further, the mindset variable (MIND), which differentiates between a deliberative and an implemental mindset, has an importance score of 1.1. Despite the positive coefficient of 1.290 suggesting that a deliberative mindset might increase overconfidence, the p-value of 0.135 indicates this effect is not statistically significant.

Similarly, the current university cycle (CYL) also shows moderate importance with a score of 0.64. The negative coefficient of -0.582 implies that as students progress from bachelor's to master's to PhD, overconfidence decreases Yet, this effect is not statistically significant with a p-value of 0.596. The expected number of tables correctly counted in the 2nd round, denoted as E[2], which also has moderate importance in the model, has a coefficient of 0.982. Still, while higher expectations in this round might theoretically increase overconfidence, the lack of statistical significance means this effect is not reliable.

4.9. Bayesian Network

In my second approach, I predicted under/overconfidence by employing a simplified Bayesian Network model. A Bayesian Network has two main advantages: (1) it is a graphical structure (with nodes and directed arcs) that represent variables and their dependencies (Jones et al., 2015); (2) incorporates prior knowledge into the model, allowing it to be refined as more data is available. Just as in Tong et al. (2021), I only used a small portion of the data to make full use of the predictors and to end with a smaller error. Therefore, the network consists of seven nodes[12] and six directed arcs, with no undirected arcs. The central node, and dependent variable, avg_delta is directly influenced by the other nodes, which are: E[2], A[1], A[2], A[3], MIND, and CYL[13]. It was assumed that the relationships between the variables follow Gaussian distributions. Hatoum et al. (2022) had already used this tool to explore the dependence relationships

[12] Once again, I applied the same approach to the best model with the interaction term. Yet, as the variables with relative importance were the same, the Bayesian Model turned out being equal.

[13] Those variables were selected because they were the ones above the average in Figure 17.

related to probability analysis and uncertainty, which also occurs in the context of my experiment.

The structure of the Bayesian network is represented in Figure 18. It illustrates that avg_delta is influenced by the aforementioned nodes, establishing a network of dependencies that are fundamental to understanding the connections within the data. Because there are no undirected arcs, the average Markov blanket size of 7.00 is equal to the number of predictors. The average neighbourhood size is 1.75, meaning that, on average, each node is directly connected to about 1.75 other nodes. This lower value compared to the Markov blanket size suggests that even though nodes have few direct connections, the overall network maintains complexity via indirect dependencies. Lastly, the average branching factor of 0.88 represents that while some nodes may influence several others, the overall tendency is towards fewer direct influences per node.

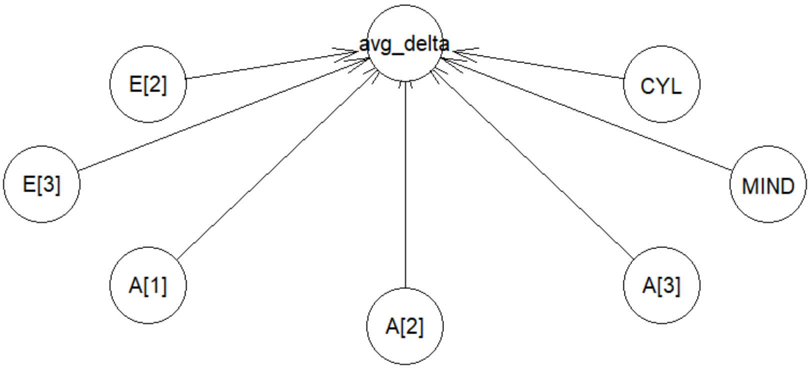


Figure 18 – Custom Bayesian network structure for predicting avg_delta.

The accuracy of the predictions was evaluated using the Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) metrics. The MSE for the model was 0.140, and the RMSE was 0.375. These error metrics point to the fact that the model has a moderate level of prediction accuracy, with the RMSE indicating that, on average, the predictions deviate from the actual values by approximately 0.375 units. Figure 19 is a proof of the goodness-of-fit of the model. The diagonal blue line in the plot represents the line of perfect prediction, where the predicted values match the observed values exactly. Each point represents a pair of observed and predicted values. The closer these points are to the diagonal blue line, the better the model's predictions align with the actual values. Here, there is a concentration of points around the diagonal line, indicating that the model's predictions are generally close to the observed values. However, there are some deviations, particularly at the extremes of the predicted values, which suggest areas where the model's accuracy could be improved. The plot supports the RMSE value by visually demonstrating that most predictions fall near the observed values, but some outliers contribute to the overall prediction error.

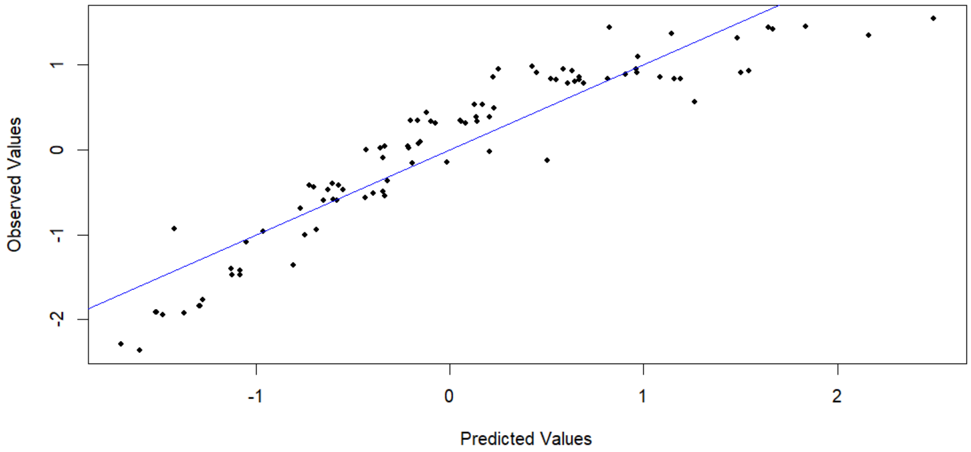


Figure 19 – Relationship between the observed values and the predicted values generated by the Bayesian Network.

Figure 20 arrives at similar conclusions regarding the outliers because the presence of residuals that are far from zero indicates that there are moments where the model's predictions are less accurate. Once again, the non-random spread of residuals seems to suggest heteroscedasticity. Regarding predictions there are two trends: (1) for predicted values in (-∞;-1] ∪ [-1;+∞), the residuals are also below the zero line, indicating that the model tends to overpredict; (2) in the interval (-1;1) the residuals are more symmetrically distributed around the zero line, indicating a more balanced prediction error. Still, given the scale of the y-axis, which ranges from -1 to 0.5, these deviations that are observed are relatively small.

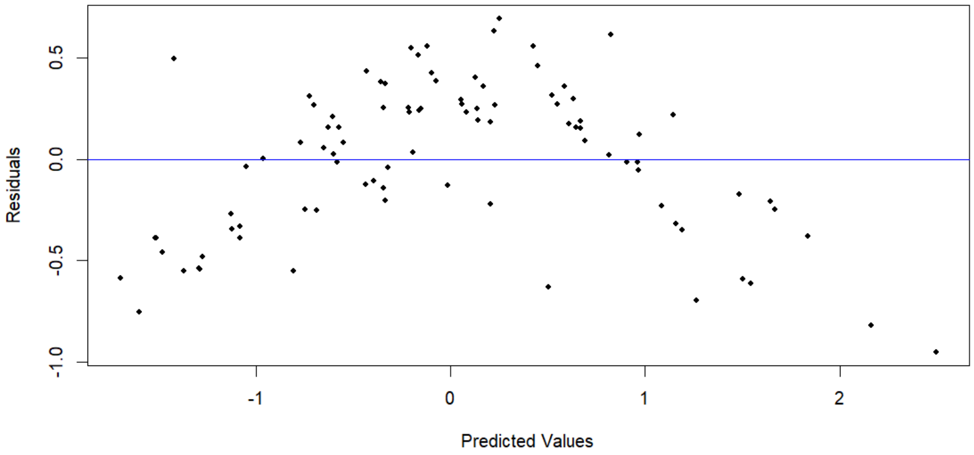


Figure 20 – Residuals vs Predicted Plot from the Bayesian Network.

Then, I replicated the same graph as in regression model 4 to quantify how much each predictor contributes to predicting the target variable. The figure below reveals that A[1], A[2] and A[3] are the most significant predictors for avg_delta, with importance values well above average. Other predictors, such as CYL, E[2] and MIND and, have considerably lower significance values, indicating that, when considering the entire network structure and the conditional dependencies, they contribute the less to the prediction of under/overconfidence.

To visualize the relationship between avg_delta and its predictors, Figure 22 presents positive correlations in red, while negative correlations appear in blue. The strong negative correlations of A[1] (-0.74), A[2] (-0.72) and A[3] (-0.69) with the target variable are consistent with their high importance values in Figure 21. Looking at the case of E[2], in the matrix it exhibits a moderate negative correlation (-0.48). However, in the Bayesian Network, E[2] only has a relative importance of 0.08. This discrepancy indicates that the predictive power of E[2] is overshadowed by other stronger predictors like A[1], A[2], and A[3]. These predictors collectively capture most of the variability in the target variable, making the additional contribution of E[2] minimal in the context of the full model.

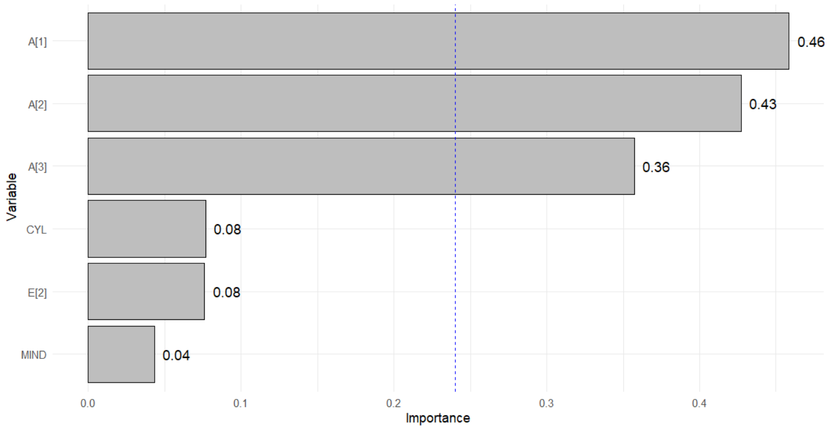
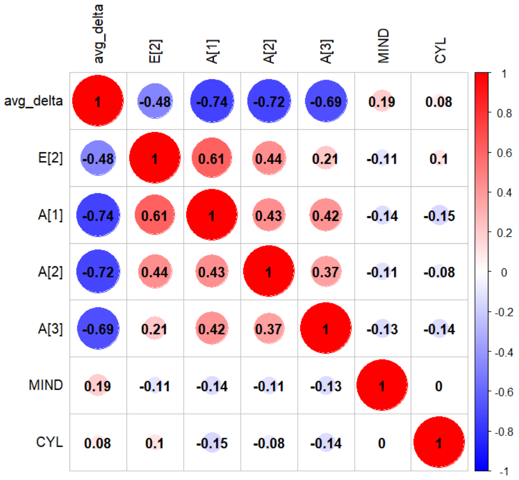


Figure 21 – Relative importance of the variables in the Bayesian network.



	avg_delta	E[2]	A[1]	A[2]	A[3]	MIND	CYL
avg_delta	1	-0.48	-0.74	-0.72	-0.69	0.19	0.08
E[2]	-0.48	1	0.61	0.44	0.21	-0.11	0.1
A[1]	-0.74	0.61	1	0.43	0.42	-0.14	-0.15
A[2]	-0.72	0.44	0.43	1	0.37	-0.11	-0.08
A[3]	-0.69	0.21	0.42	0.37	1	-0.13	-0.14
MIND	0.19	-0.11	-0.14	-0.11	-0.13	1	0
CYL	0.08	0.1	-0.15	-0.08	-0.14	0	1

Figure 22 – Correlation matrix for the Bayesian network.

Figure 23 reveals the shape, spread, skewness, and modality of each variable, providing information on the data structure. Firstly, variables A[2], A[3] and E[2] show multimodal distributions, indicating the presence of subgroups of subjects that have different performance or expectation levels. On the contrary, the symmetric and unimodal distribution of A[1] suggests a balanced performance level, with most scores clustering around a central value. The U-shaped distribution of MIND reflects its binary categorization, showing that participants could either be classified as being in an implemental or deliberative mindset. Then, the bimodal distribution of CYL represents the dominance of two main educational levels in the sample – bachelor's and master's degree, respectively. Finally, for avg_delta most values are clustered around the mean, confirming the moderate levels of under/overconfidence for the majority of cases. The right skewness, confirmed by the Shapiro-Wilk test (p = 4.68e-04 < 0.001), implies that there are a few instances where overconfidence is quite high.

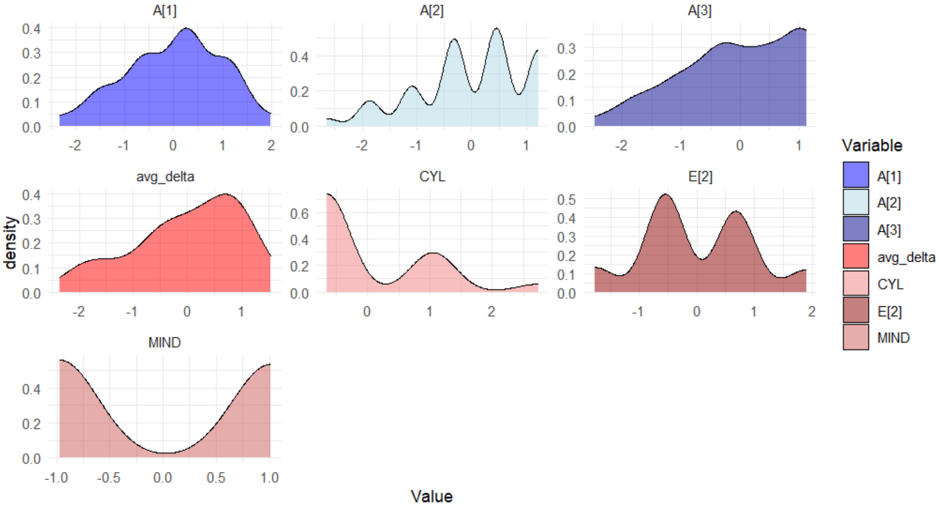


Figure 23 – Density plots for the Bayesian network.

In the end, I decided to take a chance and build an alternative Bayesian network model using only the most important predictors: A[1], A[2] and A[3]. The performance of this even more refined model was tested, obtaining an MSE of 0.150 (compared to 0.140 for the simplified model) and an RMSE of 0.387 (compared to 0.375 for the simplified model). These metrics, although similar to those obtained with the simplified model, indicate that excluding the least important variables worsened the accuracy of the forecast.

5. CONCLUSION

In this work, the main variable of study was under/overconfidence among university students, specifically how it is influenced by information about the performance of previous cohorts. I implemented a lab experiment to investigate this issue, using a between-subjects design and dividing the participants into four treatments. At the time of the predictions, the control groups (T1 and T2) did not have access to information, while the treatment groups (T3 and T4) received clear and visual information, including a graph with the distribution of past cohort performance.

The experiment followed a methodology derived from Abeler et al. (2011), applying a similar real-effort task that involved counting the number of zeros in tables. This task was chosen because it is straightforward and does not require specific prior knowledge, which makes it ideal for assessing the accuracy of predictions in a controlled environment. The relationship between psychological variables and prediction accuracy was also explored through specific questions designed to measure mindset and self-esteem. Participants answered three dichotomous questions related to their performance, which allowed them to categorise their mindset as deliberative or implemental. In addition, they answered a 10-item questionnaire to measure their position in terms of self-esteem on the Rosenberg Self-Esteem Scale.

The study faced some limitations. Running the experiment on Qualtrics and using Excel for the payments was a logistical challenge. Furthermore, the system's inability to update itself continuously meant that the payment scheme could not be based on predictions of relative positioning. Asking for percentiles for relative overconfidence led several participants to have the same percentile, due to the low variability of the options, so I had to divide them into quartiles *ex-post*. This would be a methodological improvement to apply if the study were to be replicated.

The hypotheses tested in this experiment were limited to the specific context of the study and therefore lacked external validity. Most of them were partially accepted, which suggests that the difficulty of the rounds may have been a confounding factor throughout the analysis. Nevertheless, the main result of this study was captured in Hypothesis 1. It showed that providing information about the performance of previous cohorts has immediate and positive effects on reducing forecasting errors among university students. This outcome highlights the potential benefits of such interventions in educational contexts for the accuracy of students' predictions of their grades, thus adapting their study and even improving academic performance.

The regression analysis began with an initial model in which variables not directly associated with performance explained 18.6 % of the variability in under/overconfidence. Model 4, which included 16 predictors, was selected as the final model due to its superior fit and lower error metrics. This model explained 86.6% of the variability in under/overconfidence. To assess the performance of the regression model, diagnostic graphs and analyses of the importance of the variables were performed. Although the residuals were normally distributed, the graphs indicated areas that needed improvement, such as dealing with non-linearity, heteroscedasticity and outliers.

A Bayesian network model was also used to predict under/overconfidence, revealing that actual performance in table counting was crucial. The model's accuracy metrics indicated moderate prediction accuracy. However, systematic patterns in the residuals suggested again potential areas to investigate. Prospective corrective actions could include examining the model for bias, incorporating additional predictors or considering non-linear modelling techniques.

Future research could explore several areas. Firstly, multiple linear regression models with categorical predictors for COND and an interaction term for CYL x YEAR should be created, to include coefficients for each level of COND (e.g., COND_2, COND_3, COND_4), which represent the difference in avg_delta compared to the baseline category (COND = 1). These would not include variables related to relative forecasts, in order to eliminate variables that distort the scale of some graphs (e.g., Figure 14). Implementing a Random Forest model could provide a more robust analysis, dealing better with non-linearity and interactions between variables than a Bayesian network.

Running regression models for each condition separately, while a great asset, would require a larger sample size to guarantee reliable and generalisable results. Investigating different types of feedback and their impact on prediction accuracy and self-esteem would provide further insight into the effectiveness of different educational interventions. In addition, further calculations could explore the interactions between mindset, self-esteem and demographics, involving more detailed subgroup analyses and sophisticated statistical modelling to uncover deeper information about how these factors influence prediction accuracy and under/overconfidence.

Based on this knowledge, I suggest that universities, probably starting with ISEG, conduct randomised controlled trials (RCTs) in which professors do the following: (1) provide distributions of previous grades; (2) allow students to predict their grades before tests; (3) give feedback and repeat the process. These low-cost studies would not only improve the accuracy of student predictions but also test the validity of these results in different contexts and with larger samples.

6. REFERENCES

Abeler, J., Falk, A., Goette, L., & Huffman, D. (2011). Reference Points and Effort Provision. *American Economic Review*, *101*(2), 470–492. https://doi.org/10.1257/aer.101.2.470

Acee, T. W., Hoff, M. A., Flaggs, D. A., & Sylvester, B. (2020). Time Perspective and Grade Expectations as Predictors of Student Achievement and Retention in the First Year of Community College. *Journal of College Student Retention: Research, Theory & Practice*, *24*(4), 152102512096067. https://doi.org/10.1177/1521025120960676

Armstrong, M. J., & MacKenzie, H. F. (Herb). (2017). Influence of anticipated and actual grades on studying intentions. *The International Journal of Management Education*, *15*(1), 49–59. https://doi.org/10.1016/j.ijme.2017.01.003

Barreda-Tarrazona, I., García-Gallego, A., García-Segarra, J., & Ritschel, A. (2022). A gender bias in reporting expected ranks when performance feedback is at stake. *Journal of Economic Psychology*, *90*, 102505. https://doi.org/10.1016/j.joep.2022.102505

Basse, G., Ding, P., Feller, A., & Toulis, P. (2024). Randomization Tests for Peer Effects in Group Formation Experiments. *Econometrica*, *92*(2), 567–590. https://doi.org/10.3982/ecta20134

Bayer, U. C., & Gollwitzer, P. M. (2005). Mindset effects on information search in self-evaluation. *European Journal of Social Psychology*, *35*(3), 313–327. https://doi.org/10.1002/ejsp.247

Brandstätter, V., Giesinger, L., Job, V., & Frank, E. (2015). The Role of Deliberative Versus Implemental Mindsets in Time Prediction and Task Accomplishment. *Social Psychology*, *46*(2), 104–115. https://doi.org/10.1027/1864-9335/a000231

Caplan, D., Mortenson, K. G., & Lester, M. (2017). Can incentives mitigate student overconfidence at grade forecasts? *Accounting Education*, *27*(1), 27–47. https://doi.org/10.1080/09639284.2017.1361850

Chapkovski, P., & Kujansuu, E. (2019). Real-time interactions in oTree using Django Channels: Auctions and real effort tasks. *Journal of Behavioral and Experimental Finance*, *23*, 114–123. https://doi.org/10.1016/j.jbef.2019.05.008

Charalambous, Y., & Charalambous, C. Y. (2023). Examining the effects of an intervention on mathematical modeling in problem solving at upper elementary grades: a cluster randomized trial study. *Mathematical Thinking and Learning*, 1–19. https://doi.org/10.1080/10986065.2023.2270088

Charness, G., Gneezy, U., & Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, *131*, 141–150. https://doi.org/10.1016/j.jebo.2016.08.010

Charness, G., Gneezy, U., & Henderson, A. (2018). Experimental methods: Measuring effort in economics experiments. *Journal of Economic Behavior & Organization*, *149*, 74–87. https://doi.org/10.1016/j.jebo.2018.02.024

Conner, J. O. (2022). Applying experiential learning theory to student activism. *Journal of Further and Higher Education*, *46*(9), 1–14. https://doi.org/10.1080/0309877x.2022.2061843

Didier Demazière. (2024). In search of a job—But which one? How unemployed people revise their occupational expectations. *Social Policy & Administration/Social Policy and Administration*. https://doi.org/10.1111/spol.13011

Ehrlinger, J., Johnson, K., Banner, M., Dunning, D., & Kruger, J. (2008). Why the unskilled are unaware: Further explorations of (absent) self-insight among the incompetent. *Organizational Behavior and Human Decision Processes*, *105*(1), 98–121. https://doi.org/10.1016/j.obhdp.2007.05.002

Erat, S., Demirkol, K., & Sallabas, M. E. (2020). Overconfidence and its link with feedback. *Active Learning in Higher Education*, *23*(3), 146978742098173. https://doi.org/10.1177/1469787420981731

Geraci, L., Kurpad, N., Tirso, R., Gray, K. N., & Wang, Y. (2022). Metacognitive errors in the classroom: The role of variability of past performance on exam prediction accuracy. *Metacognition and Learning*. https://doi.org/10.1007/s11409-022-09326-7

Gollwitzer, P. M., Heckhausen, H., & Steller, B. (1990). Deliberative and implemental mind-sets: Cognitive tuning toward congruous thoughts and information. *Journal of Personality and Social Psychology*, *59*(6), 1119–1127. https://doi.org/10.1037/0022-3514.59.6.1119

Gric, Z., Ehrenbergerova, D., & Hodula, M. (2022). The power of sentiment: Irrational beliefs of households and consumer loan dynamics. *Journal of Financial Stability*, *59*, 100973. https://doi.org/10.1016/j.jfs.2022.100973

Guskey, T. R. (2022). Can grades be an effective form of feedback? *Phi Delta Kappan*, *104*(3), 36–41. https://doi.org/10.1177/00317217221136597

Hamann, K., Pilotti, M. A. E., & Wilson, B. M. (2020). Students' Self-Efficacy, Causal Attribution Habits and Test Grades. *Education Sciences*, *10*(9), 231. https://doi.org/10.3390/educsci10090231

Han, Z. (2020). Low-frequency fiscal uncertainty. *Journal of Monetary Economics*, *117*. https://doi.org/10.1016/j.jmoneco.2020.03.017

Hatoum, K., Moussu, C., & Gillet, R. (2022). CEO overconfidence: Towards a new measure. *International Review of Financial Analysis*, *84*, 102367. https://doi.org/10.1016/j.irfa.2022.102367

Hossain, B., & Tsigaris, P. (2012). Are grade expectations rational? A classroom experiment. *Education Economics*, *23*(2), 199–212. https://doi.org/10.1080/09645292.2012.735073

Hügelschäfer, S., & Achtziger, A. (2014). On confident men and rational women: It's all on your mind(set). *Journal of Economic Psychology*, *41*, 31–44. https://doi.org/10.1016/j.joep.2013.04.001

Jiménez-Jiménez, N., Molis, E., & Solano-García, Á. (2023). Don't shoot yourself in the foot! A (real-effort task) experiment on income redistribution and voting. *European Journal of Political Economy*, *78*, 102325. https://doi.org/10.1016/j.ejpoleco.2022.102325

Jones, A. M., Lomas, D., Moore, P., & Rice, N. (2015). A Quasi-Monte-Carlo Comparison of Parametric and Semiparametric Regression Methods for Heavy-tailed and Non-normal Data: an Application to Healthcare Costs. *Journal of the Royal Statistical Society*, *179*(4), 951–974. https://doi.org/10.1111/rssa.12141

Kahneman, D. (2011). *Thinking, fast and slow*. Farrar, Straus and Giroux.

Karaca, M., Geraci, L., Kurpad, N., Lithander, M. P. G., & Balsis, S. (2023). Low-Performing Students Confidently Overpredict Their Grade Performance throughout the Semester. *Journal of Intelligence*, *11*(10), 188. https://doi.org/10.3390/jintelligence11100188

Keller, L., & Gollwitzer, P. M. (2017). Mindsets Affect Risk Perception and Risk-Taking Behavior. *Social Psychology*, *48*(3), 135–147. https://doi.org/10.1027/1864-9335/a000304

Khattab, N., Madeeha, M., Samara, M., Modood, T., & Barham, A. (2021). Do educational aspirations and expectations matter in improving school achievement? *Social Psychology of Education*, *25*(1), 33–53. https://doi.org/10.1007/s11218-021-09670-7

Koenka, A. C. (2020). Grade expectations: The motivational consequences of performance feedback on a summative assessment. *The Journal of Experimental Education*, *90*(1), 1–24. https://doi.org/10.1080/00220973.2020.1777069

Kolovelonis, A., & Goudas, M. (2018). The relation of physical self-perceptions of competence, goal orientation, and optimism with students' performance calibration in physical education. *Learning and Individual Differences*, *61*, 77–86. https://doi.org/10.1016/j.lindif.2017.11.013

Li, J., Hügelschäfer, S., & Achtziger, A. (2018). A self-regulatory approach to rational decisions: The implemental mindset optimizes economic decision making in situations requiring belief updating. *Journal of Theoretical Social Psychology*, *3*(2), 115–126. https://doi.org/10.1002/jts5.38

Magnus, J. R., & Peresetsky, A. A. (2018). Grade Expectations: Rationality and Overconfidence. *Frontiers in Psychology*, *8*. https://doi.org/10.3389/fpsyg.2017.02346

Nederhand, M. L., Tabbers, H. K., Jongerling, J., & Rikers, R. M. J. P. (2020). Reflection on exam grades to improve calibration of secondary school students: a longitudinal study. *Metacognition and Learning*. https://doi.org/10.1007/s11409-020-09233-9

O'Reilly, C. A., & Hall, N. (2021). Grandiose narcissists and decision making: Impulsive, overconfident, and skeptical of experts–but seldom in doubt. *Personality and*

Individual Differences, *168*(110280), 110280. https://doi.org/10.1016/j.paid.2020.110280

OECD. (2023). Education at a Glance 2023. In *Education at a glance*. OECD. https://doi.org/10.1787/e13bef63-en

Pennings, M., Cushing, D. F., Gomez, R., Dyson, C., & Coombs, C. (2019). Gaining "Raw Insider Knowledge": The Benefits and Challenges of International Study Tours for Creative Industries Students. *International Journal of Art & Design Education*, *38*(2), 539–554. https://doi.org/10.1111/jade.12219

Rodrigo-González, A., Caballer-Tarazona, M., & García-Gallego, A. (2021). Effects of Inequality on Trust and Reciprocity: An Experiment With Real Effort. *Frontiers in Psychology*, *12*. https://doi.org/10.3389/fpsyg.2021.745948

Rosenberg, M. (1965). Rosenberg Self-Esteem Scale. *PsycTESTS Dataset*, *1*(1). https://doi.org/10.1037/t01038-000

Ruthig, J. C., & Kroke, A. M. (2024). Improving accuracy in predictions about future performance. *Current Psychology*. https://doi.org/10.1007/s12144-024-06023-3

Sabater-Grande, G., Georgantzís, N., & Herranz-Zarzoso, N. (2022). Goals and guesses as reference points: a field experiment on student performance. *Theory and Decision*. https://doi.org/10.1007/s11238-022-09892-x

Saenz, G. D., Geraci, L., & Tirso, R. (2019). Improving metacognition: A comparison of interventions. *Applied Cognitive Psychology*, *33*(5). https://doi.org/10.1002/acp.3556

Sagar, V., Sengupta, R., & Sridharan, D. (2019). Dissociable sensitivity and bias mechanisms mediate behavioral effects of exogenous attention. *Scientific Reports*, *9*(1). https://doi.org/10.1038/s41598-019-42759-w

Silva, A. D., Vautero, J., & Usssene, C. (2021). The influence of family on academic performance of Mozambican university students. *International Journal of Educational Development*, *87*, 102476. https://doi.org/10.1016/j.ijedudev.2021.102476

Snyder, K. E., Barr, S. M., Honken, N. B., Pittard, C. M., & Ralston, P. A. S. (2018). Navigating the First Semester: An Exploration of Short-Term Changes in

Motivational Beliefs Among Engineering Undergraduates. *Journal of Engineering Education*, *107*(1), 11–29. https://doi.org/10.1002/jee.20187

Subramaniam, K. (2022). A phenomenological study of prospective teachers' first-time science teaching experiences. *Teaching Education*, *34*(2), 1–16. https://doi.org/10.1080/10476210.2022.2077928

Tirso, R., Geraci, L., & Saenz, G. D. (2019). Examining Underconfidence Among High-Performing Students: A Test of the False Consensus Hypothesis. *Journal of Applied Research in Memory and Cognition*, *8*(2), 154–165. https://doi.org/10.1016/j.jarmac.2019.04.003

Tong, L.-L., Gu, J.-B., Li, J.-J., Liu, G.-X., Jin, S.-W., & Yan, A.-Y. (2021). Application of Bayesian network and regression method in treatment cost prediction. *BMC Medical Informatics and Decision Making*, *21*(1). https://doi.org/10.1186/s12911-021-01647-y

Vignery, K. (2022). From networked students centrality to student networks density: What really matters for student performance? *Social Networks*, *70*, 166–186. https://doi.org/10.1016/j.socnet.2022.01.001

Wright, N. A., & Arora, P. (2022). A for effort: Incomplete information and college students' academic performance. *Economics of Education Review*, *88*, 102238. https://doi.org/10.1016/j.econedurev.2022.102238

Wyness, G., Macmillan, L., Anders, J., & Dilnot, C. (2022). Grade expectations: how well can past performance predict future grades? *Education Economics*, *31*(4), 1–22. https://doi.org/10.1080/09645292.2022.2113861

Yang, Q., Ybarra, O., Van den Bos, K., Zhao, Y., Guan, L., Cao, Y., Li, F., & Huang, X. (2019). Neurophysiological and behavioral evidence that self-uncertainty salience increases self-esteem striving. *Biological Psychology*, *143*, 62–73. https://doi.org/10.1016/j.biopsycho.2019.02.011

APPENDICES

Appendix A – Wilcoxon Rank-Sum test results for A[n] and E[n]

TABLE VII

A[N] COMPARISONS

	A[1]	A[2]	A[3]
T1 vs T2	0.666	0.063	0.000
T3 vs T4	0.964	0.001	0.107

TABLE VIII

E[N] COMPARISONS

	E[1]	E[2]	E[3]
T1 vs T2	0.090	0.325	0.103
T3 vs T4	0.283	0.143	0.512

Appendix B – Pre-test Link and Instructions

Qualtrics link

https://ucpresearch.qualtrics.com/jfe/form/SV_9YNJ00vAkGwLH1Q

Detailed Instructions

Instructions (Page 1 of 2)

The experiment consists of 3 rounds that last 120 seconds (2 minutes) each. Please read the explanations carefully.

If you have any doubts or issues while completing the experiment, please call the researcher/s present in the room by raising your hand.

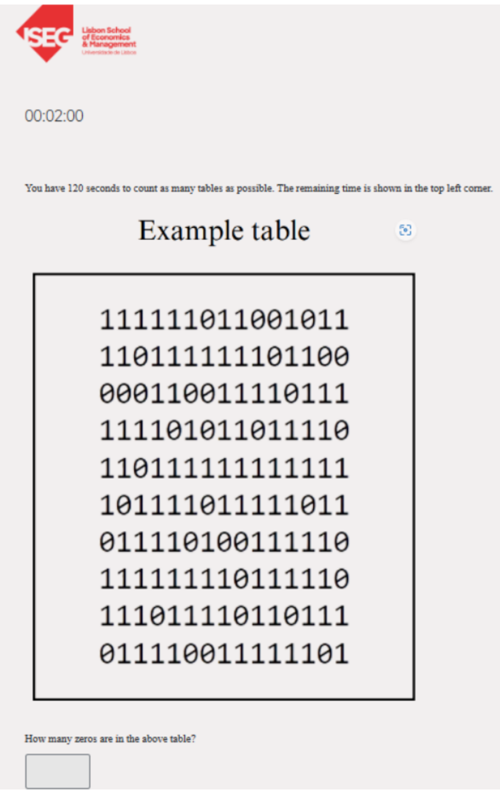
Over the experiment, there will be **no repeated questions**. In addition, the order of the rounds may be different, so there is no advantage in copying from the colleague next to you or memorizing past answers.

At the end, you will be asked to create an **ID** to be eligible to participate in the lottery to win one of the Gift Cards. In case you are one of the winners, your final earnings (in credits) will be converted into euros as follows: **1 credit = 1 euro**. Roundings may be made at the time of payment, but always for your benefit.

Click on the "next" button to proceed.

Instructions (Page 2 of 2)

Your task consists of counting zeros within a limited time. This figure is an example of the screen you will see later.



00:02:00

You have 120 seconds to count as many tables as possible. The remaining time is shown in the top left corner.

Example table

```
111111011001011
110111111101100
000110011110111
111101011011110
110111111111111
101111011111011
011110100111110
111111110111110
111011110110111
011110011111101
```

How many zeros are in the above table?

After counting the **number of zeros** in a table, enter that number in the box below.

During the 2 minutes you have per round, you may **count up to 5 tables** correctly. The remaining time will appear in the top left corner. When time runs out, the study will advance automatically.

Payment: For participating, you start with **4.75 credits**. You can earn more credits depending on your performance.

For each **correct** answer, you earn **0.35 credits**;

For each **wrong or blank** answer, you receive **0.00 credits** (they do not discount or add up any credits).

Thus, variable earnings can be more than fixed earnings.

Example: After the 3 rounds, you solved 7 tables correctly, 5 incorrectly and left 3 blank. Your total earnings would be:

+4.75 credits, fixed payment

+7x0.35 credits for correctly counted tables

+8x0.00 credits for incorrectly/blank counted tables

Total earnings: 4.75+7x0.35 = 7.20 credits (7.20 euros)

Note that this is just an example and should not be taken as an indication of the level of difficulty present in this experiment.

Counting tips: All strategies are valid, so you are free to count the zeros as you wish. Yet, experience shows that it helps to count the zeros in pairs and, at the end, multiply the number of zeros by two. Placing the cursor/finger over the number you are counting reduces the number of mistakes. Reporting the number of zeros immediately after counting them also reduces blank answers.

Click on the "next" button if you are ready to start the first round.

Appendix C – Experimental Links

T1: https://ucpresearch.qualtrics.com/jfe/form/SV_29V6EMM91wWDdVY

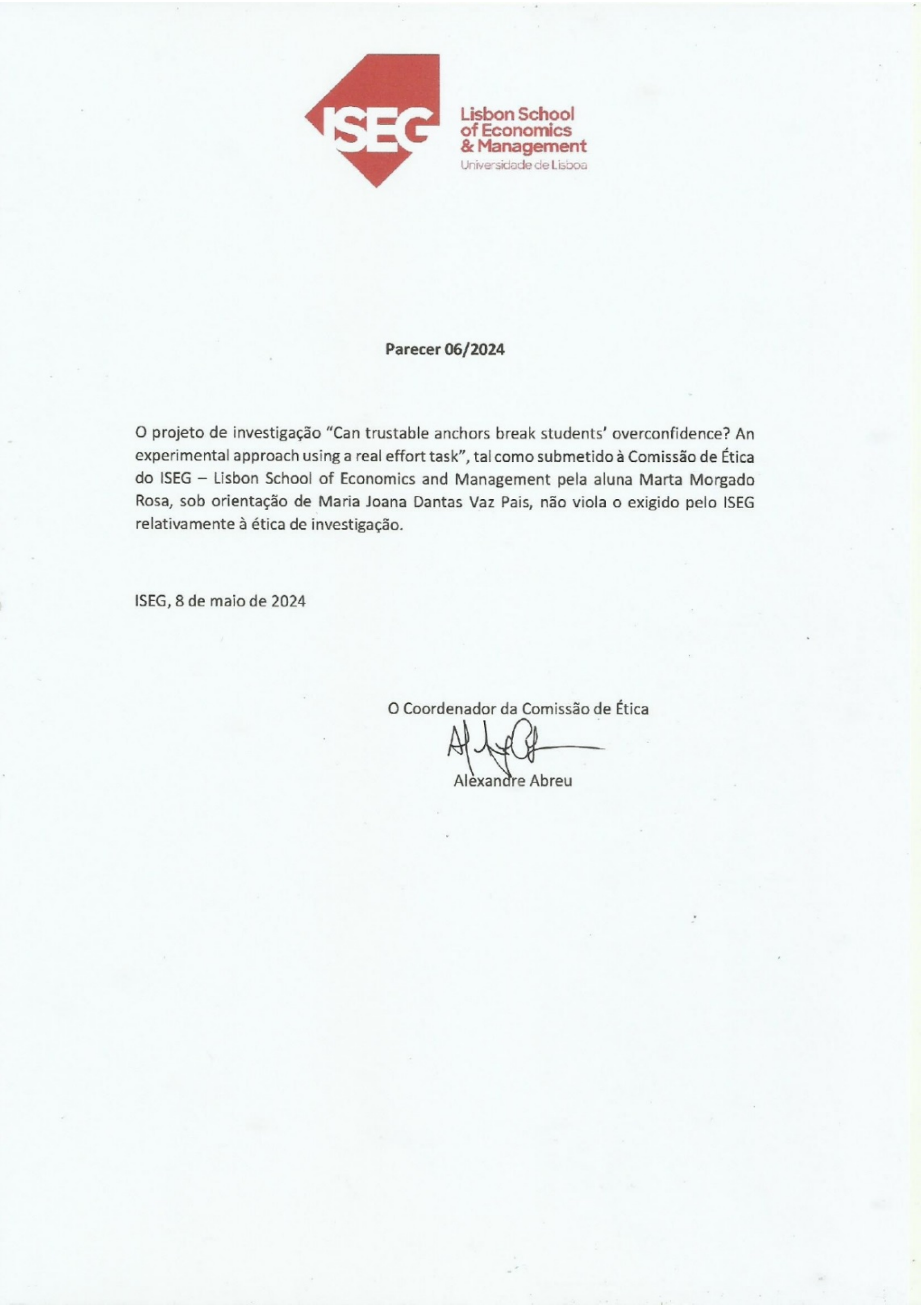
T2: https://ucpresearch.qualtrics.com/jfe/form/SV_5C2UbM0UbfEzwLY

T3: https://ucpresearch.qualtrics.com/jfe/form/SV_9XnbiCZ4cmTsfP0

T4: https://ucpresearch.qualtrics.com/jfe/form/SV_e8UNaGS4Vn7BKcu

Appendix D – Photo of the Laboratory at XLAB

Appendix E – Ethical Approval



ISEG Lisbon School of Economics & Management
Universidade de Lisboa

Parecer 06/2024

O projeto de investigação "Can trustable anchors break students' overconfidence? An experimental approach using a real effort task", tal como submetido à Comissão de Ética do ISEG – Lisbon School of Economics and Management pela aluna Marta Morgado Rosa, sob orientação de Maria Joana Dantas Vaz Pais, não viola o exigido pelo ISEG relativamente à ética de investigação.

ISEG, 8 de maio de 2024

O Coordenador da Comissão de Ética

Alexandre Abreu

Appendix F – Expected and Actual Performance Graphs

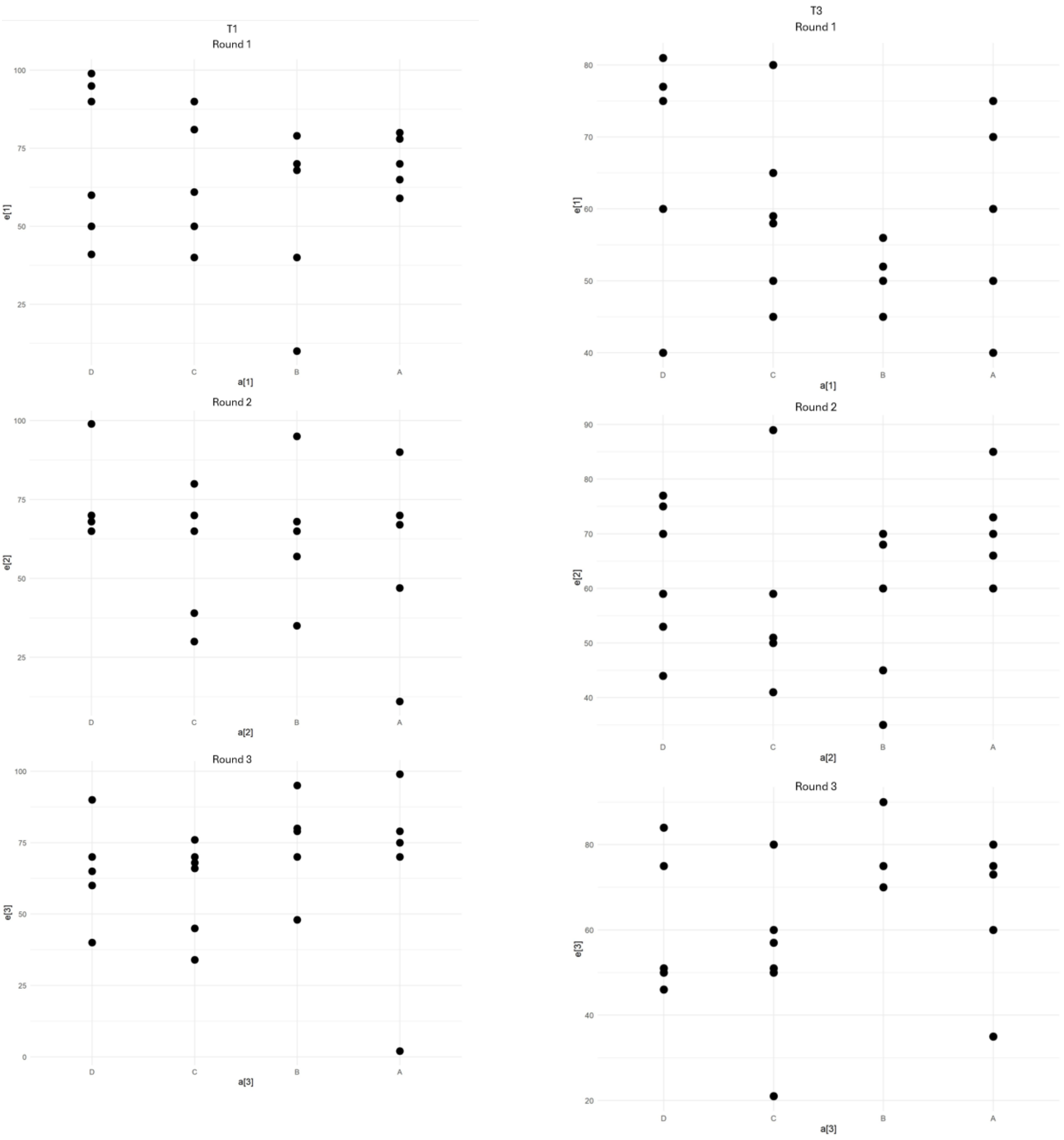


Figure 24 – On the left, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T1. On the right, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T3.

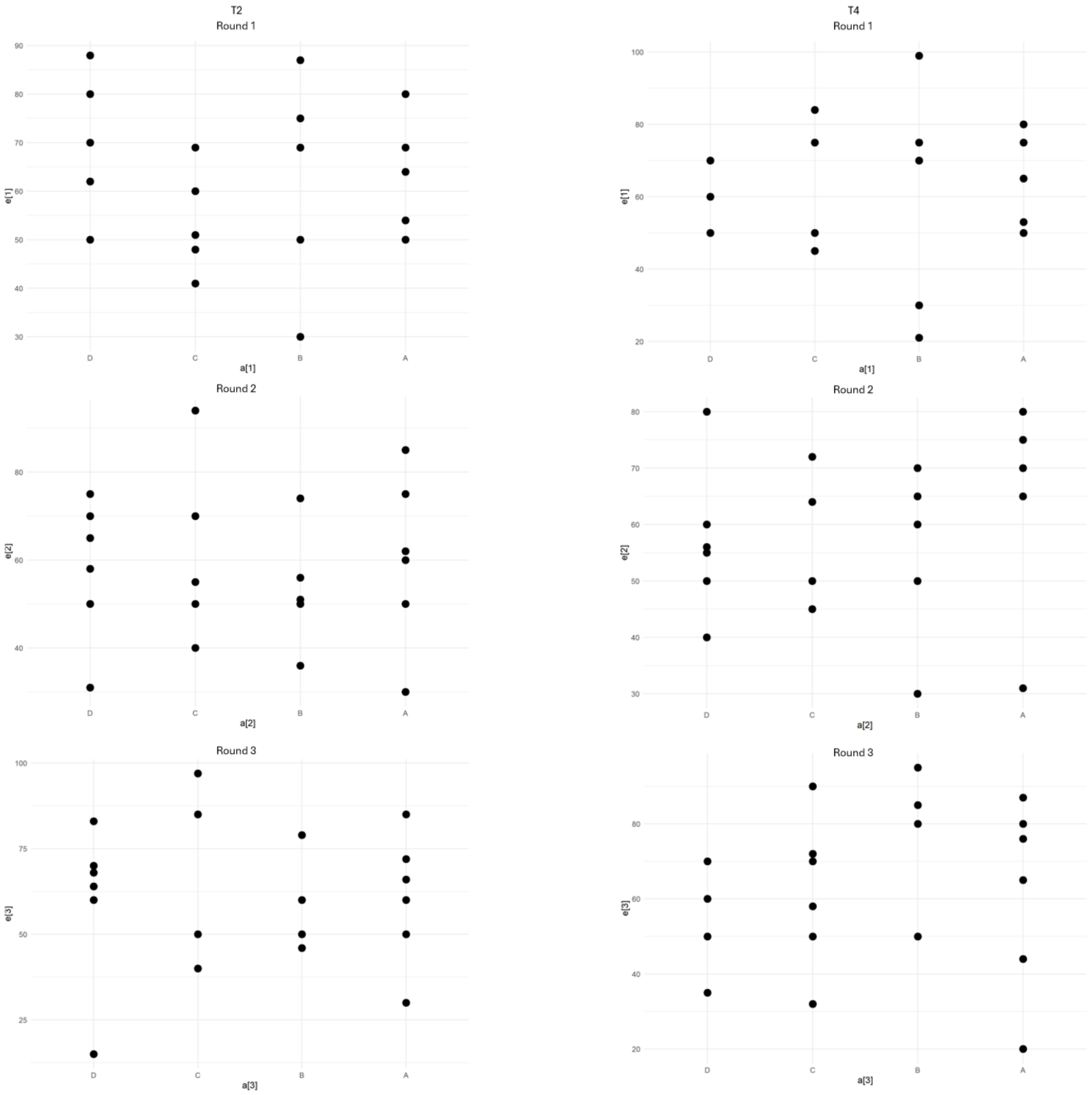


Figure 25 – On the left, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round, in T2. On the right, the relationship between the actual quartile, a[n], vs the expected percentile, e[n], relative to peers, per round in, T4.

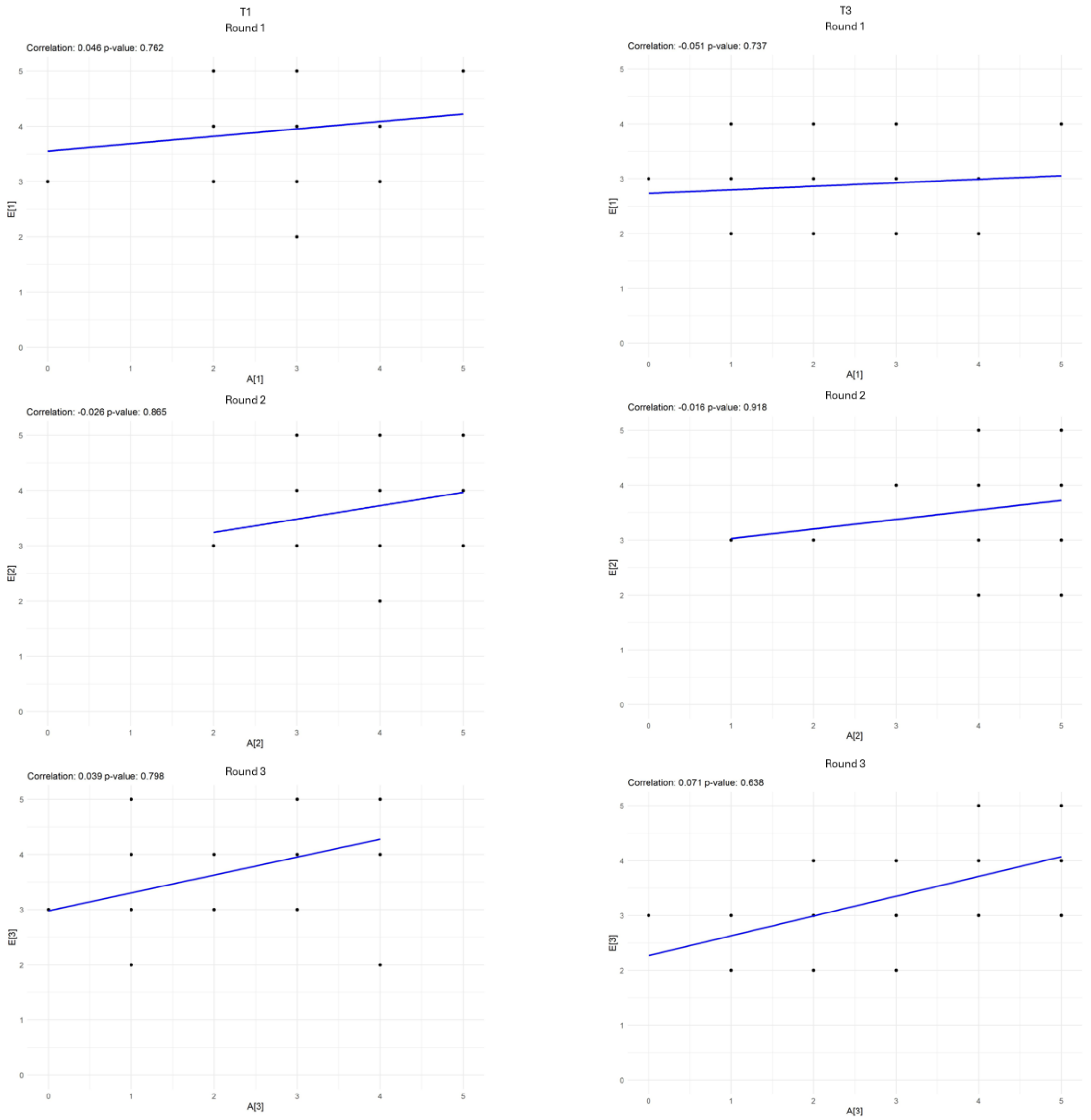


Figure 26 – On the left, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T1. On the right, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T3.

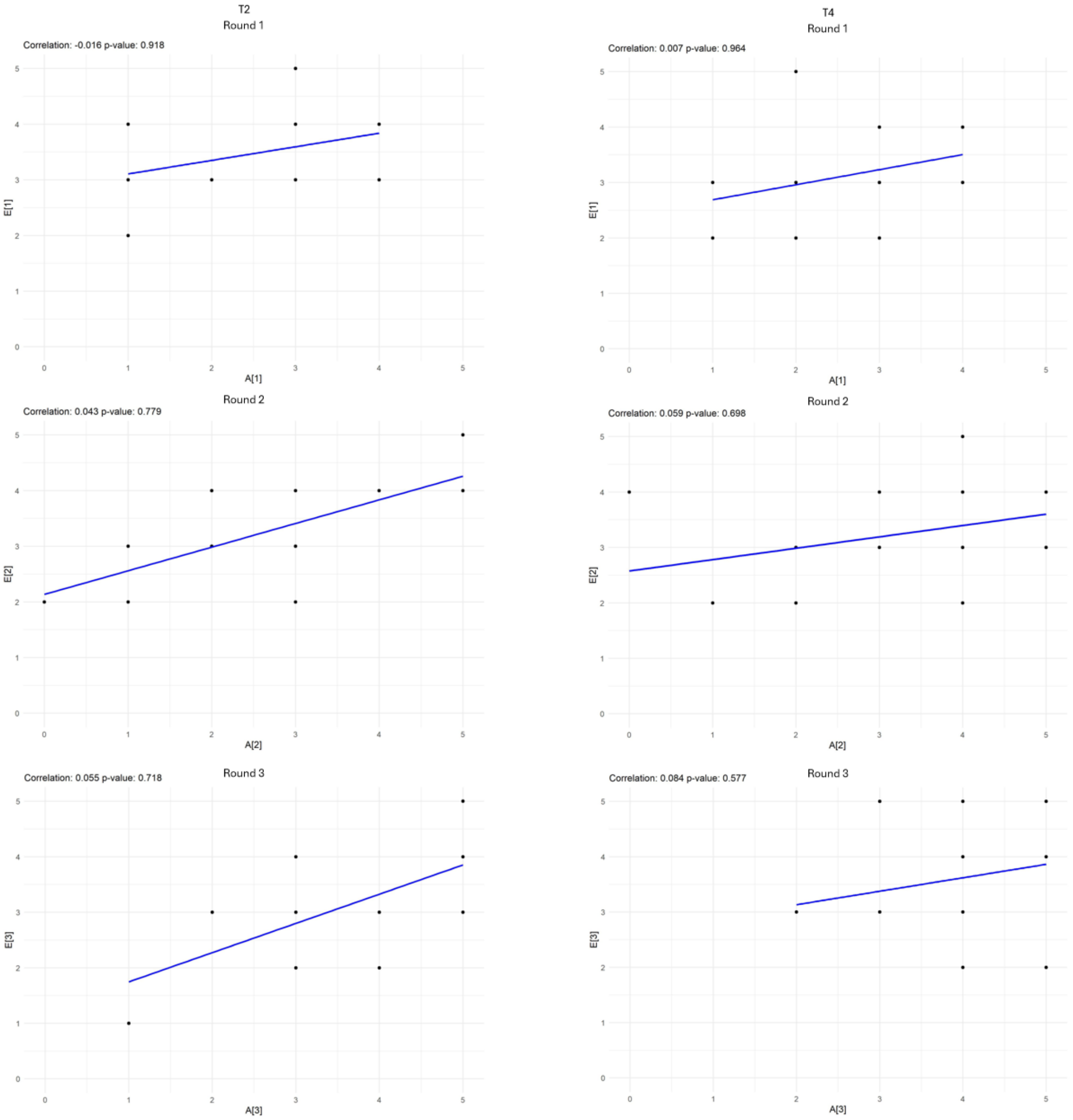


Figure 27 – On the left, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T2. On the right, the correlation between the actual, A[n], vs the expected, E[n], number of tables correctly counted before the task was known, per round, in T4.

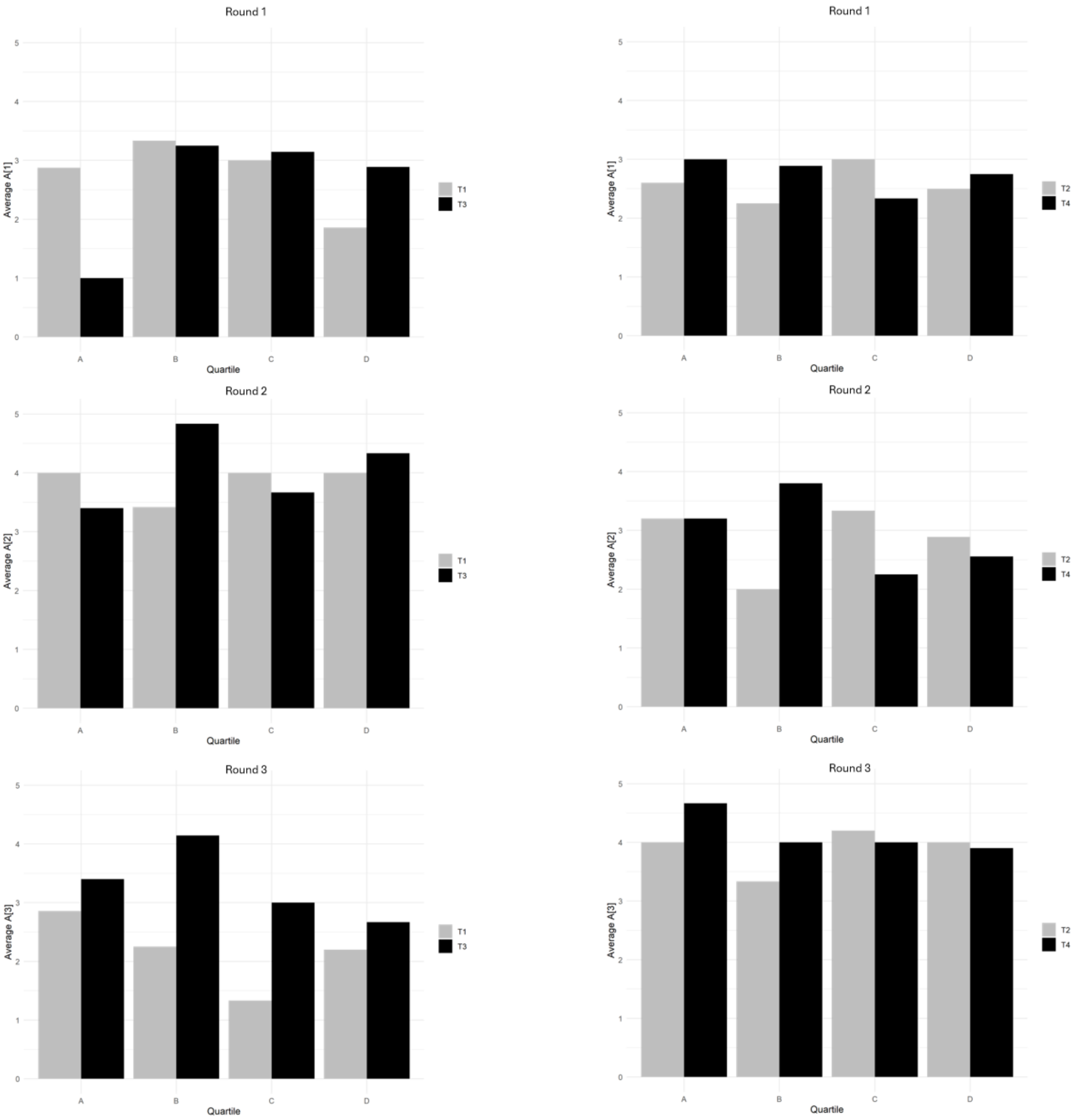


Figure 28 – On the left, the relationship between the expected percentile, e[n], converted into quartiles, relative to peers vs the average actual number of tables correctly counted, A[n], before the task was known, between T1 and T3. On the right, the relationship between the expected percentile, e[n], converted into quartiles, relative to peers vs the average actual number of tables correctly counted, A[n], before the task was known, between T2 and T4.

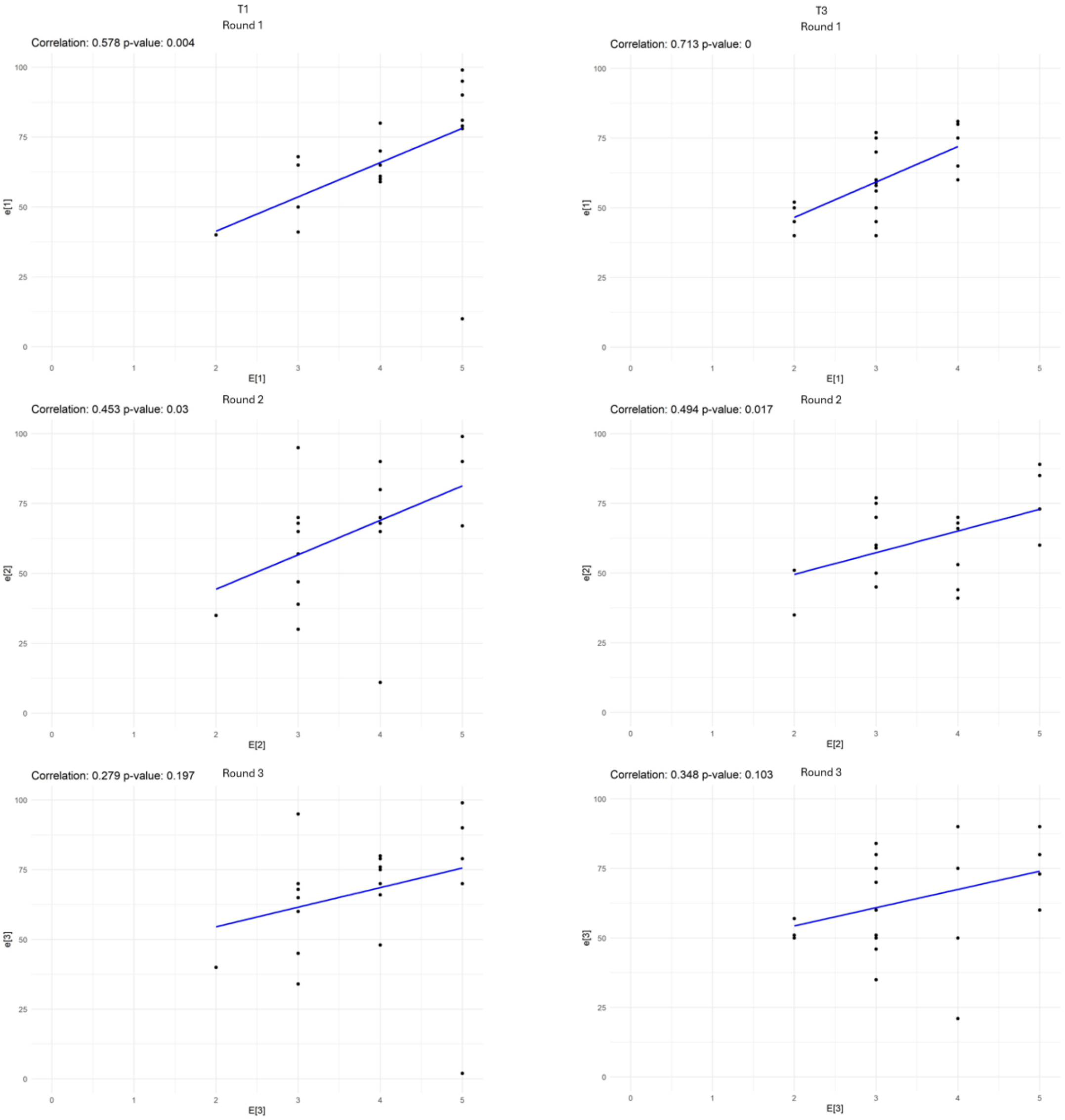


Figure 29 – On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs the expected percentile relative to peers, e[n], per round, in T1. On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs expected percentile relative to peers, e[n], per round, in T3.

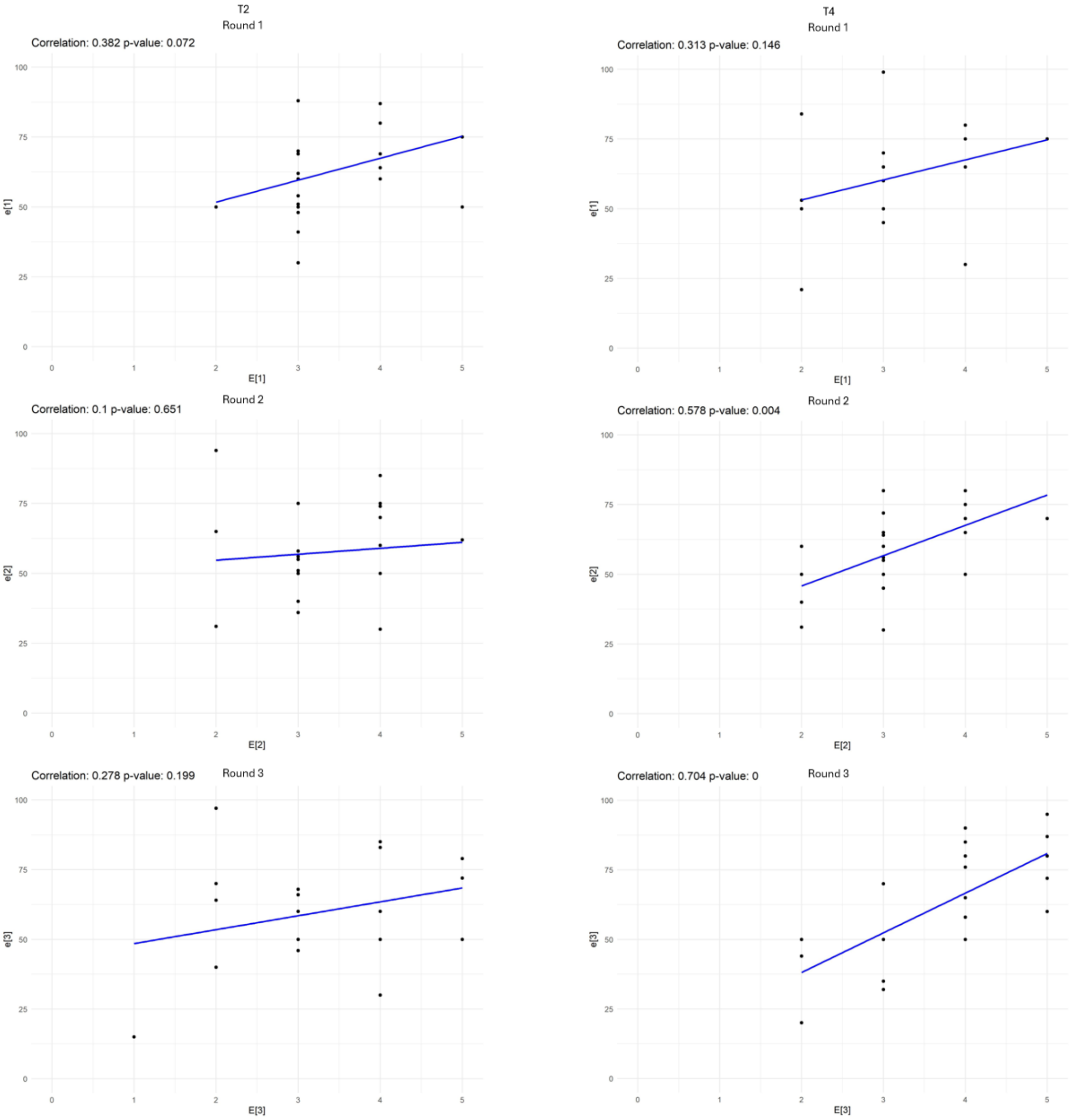


Figure 30 – On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs the expected percentile relative to peers, e[n], per round, in T2. On the left, the correlation between the expected, A[n], number of tables correctly counted before the task was known, vs expected percentile relative to peers, e[n], per round, in T4.

Appendix G – Predictors of avg_delta with the interaction term

TABLE IX

PREDICTORS OF AVG_DELTA INCLUDING CYL X AGE

	Model 1		Model 2		Model 3		Model 4	
Predictors	β_i	p-Value	β_i	p-Value	β_i	p-Value	β_i	p-Value
(Intercept)	-9.684	0.522	4.325	0.546	4.530	0.526	4.010	0.567
COND	-1.327	0.110	-0.042	0.927	-0.057	0.901	-0.050	0.912
E[1]	-	-	0.649	0.331	0.480	0.455	0.645	0.331
E[2]	-	-	0.979	0.268	0.808	0.349	1.053	0.207
E[3]	-	-	-0.685	0.301	-0.462	0.455	-0.710	0.275
e[1]	-	-	-0.056	0.240	-0.040	0.365	-0.052	0.247
e[2]	-	-	0.015	0.790	0.051	0.239	-	-
e[3]	-	-	0.046	0.342	-	-	0.054	0.135
A[1]	-	-	-3.709	7.60e-10***	-3.760	3.89e-10***	-3.193	5.50e-10***
A[2]	-	-	-3.165	2.52e-10***	-3.011	1.04e-10***	--3.193	5.37e-11**
A[3]	-	-	-2.217	4.32e-07***	-2.236	3.31e-07***	-2.214	3.72e-07***
MIND	3.350	0.082	1.249	0.161	1.108	0.206	1.294	0.137
SE	-0.382	0.052	-0.038	0.698	-0.035	0.724	-0.037	0.703
GEN	1.054	0.599	-0.133	0.886	-0.203	0.827	-0.149	0.872
AGE	-0.090	0.874	-0.146	0.586	-0.143	0.593	-0.142	0.592
CYL	12.882	0.006**	-1.317	0.583	-1.416	0.555	-1.159	0.616
YEAR	6.727	0.027*	0.110	0.941	0.034	0.982	0.215	0.881
GPA	1.436	0.011*	0.116	0.656	0.107	0.681	0.116	0.656
CYL X AGE	4.400	0.034*	0.358	0.734	0.386	0.714	0.288	0.776
Other Diagnostics								
Root MSE	8.480		3.747		3.745		3.723	
F-statistics	2.71	0.008	26.22	<2.2e-16	27.74	<2.2e-16	28.11	<2.2e-16
Multiple R^2	0.229		0.866		0.864		0.866	
Adj.R^2	0.145		0.833		0.833		0.835	

*p-Value < 0.05; **p-Value < 0.01; *** p-Value < 0.001

Appendix H – Experimental databases and codes

https://www.dropbox.com/scl/fo/4zpy4b3uo6wmapdwzw8zr/AFZjXBreEx8AoCxwZ7Yy9-c?rlkey=zbixgbcpdgwv8h4sat9qk42hn&st=p9cqmc5m&dl=0