The effect of problem solving before instruction on bayes theorem learning

CS-411
Digital Education

[redacted]

Lausanne, June 5, 2025

Contents

1 | Introduction

The traditional way of education is to first give a lecture ('instruction') and after that work on exercises ('problem solving'). In this report, we are researching the effect that reversing this traditional order has on the learning gain. To evaluate different instructional strategies, we employ two methods: PS-I (Problem Solving followed by Instruction) and I-PS (Instruction followed by Problem Solving). The PS-I approach aligns with theories of productive failure and preparation for future learning, where learners first struggle with a problem before receiving formal instruction. On the other hand, the I-PS approach follows a more traditional instructional model in which explicit teaching precedes application, providing learners with the conceptual framework before engaging in problem solving activities.

The topic on which we assess learning gain is Bayes theorem. Bayes theorem is a well-known, mathematical theorem, about conditional probability [2]. It is a rule for computing the probability that event A happens, based on the fact that we know that event B is true. We aim to teach Bayes' theorem to participants such as high school graduates, university students, or working professionals who have little to no prior exposure to probability or statistics in higher education. This target audience represents individuals who have completed secondary education but have not engaged deeply with statistical concepts in university-level coursework. These learners typically possess basic mathematical skills including arithmetic operations, fractions, and percentages, but lack formal training in probability theory or statistical reasoning.

In the PS-I condition, participants first engage in a submarine search game designed to provide intuitive exposure to the concept of Bayes' Theorem. This activity is followed by explicit instruction. Conversely, in the I-PS condition, participants receive a structured explanation of Bayes' Theorem first, and then apply their understanding through the same submarine problem-solving activity. To compare the effectiveness of these two instructional methods, we will measure learning gains by analyzing the difference between each participant's pre-test and post-test scores, allowing us to determine which instructional strategy better supports conceptual understanding and practical application of Bayes' Theorem.

In the next chapter (Section 2) the specific learning goals are presented. Then, in Section 3 the lesson design is elaborated on, including descriptions of the problem solving activity and the instruction. Section 4 highlights our experimental design, followed by the implementation of our lesson design in Section 5. The analysis of our data can be found in Section 6. Lastly, a conclusion and discussion are presented in Section 7

2 | Learning Goals

Our overall learning goal is for participants to understand the concept of Bayes' Theorem and for them to be able to apply it to real-world contexts through realistic examples. In short, we want them to recognize and solve problems that require probabilistic updates based on new evidence.

To support this goal, we define the following four measurable learning objectives:

1. **Recognition:** Distinguish situations in which Bayes' Theorem is applicable and those in which it is not. Participants must be able to differentiate between problems requiring Bayes' Theorem and those involving other types of mathematical or logical reasoning.

2. **Calculation:** Compute conditional probabilities (i.e., the posterior) using textual information that includes the prior, likelihood, and marginal probabilities, by applying Bayes' Theorem. This objective requires participants to identify the necessary components from word problems, substitute values correctly, and perform accurate calculations to determine posterior probabilities.

3. **Interpretation:** Interpret a computed posterior probability and explain its meaning in the context of a real-world scenario. This learning goal requires participants to be able to translate the numerical solution that the formula gives into a meaningful conclusion.

4. **Creation:** Formulate a real-world scenario where Bayes' Theorem would be appropriately applied. Participants will need to fully understand the concept of the theorem and also have some creativity in order to create the scenario.

These learning goals formed the basis for the design of our instructional materials, including the pre-test, the problem-solving activity, and the post-test. Each question in the pre-test and post-test will be linked to one of the four learning objectives.

3 | Lesson Design and Activities

Our lesson is logically structured to guide learners toward mastery of Bayes' Theorem. We designed the pre-test and post-test at comparable difficulty levels, aligned with the four learning goals, to accurately measure student progress. For the problem-solving activity, we carefully selected the "Find the Submarine" game, which serves as an effective example to introduce students to the intuition behind Bayes' Theorem. Regarding the instruction phase, we have carefully aligned with our learning objectives, taking into account learners' prior knowledge and cognitive development to ensure a coherent and supportive learning experience.

Pre-test

We begin with a pre-test to assess students' baseline understanding of Bayes' Theorem, conditional probability, and basic probabilistic thinking. The pre-test includes a combination of true/false questions (worth 1 point each) and single-choice questions (worth 2 points each), with a total possible score of 10 points. The assessment targets all our four learning objectives, this diagnostic step ensures we can measure learning gains and identify areas needing emphasis.

Instruction

Following the pre-test, students experience the lesson content through an instructional activity designed to be accessible and engaging. Our multimedia approach includes:

Step 1: Probability concept. We begin with a brief reading text introducing basic probability principles in case the participant forgot or does not know the basics.

Step 2: Conditional probability. A 3-minute video is provided that explores conditional probability using the classic "boy or girl" family problems, helping students see how probabilities change when additional information is provided.

Step 3: Bayes Theorem. An 8-minute video is provided that presents Bayes' Theorem using weather prediction scenarios to explain how we update beliefs based on new evidence. This real-world application makes the theorem both intuitive and memorable, demonstrating practical relevance.

These multimedia resources are selected to foster conceptual understanding by connecting abstract ideas with everyday situations. Moreover, by showing different kind of examples on conditional probability and Bayes theorem, participants are supported in reaching the first learning goal: recognition. Moreover, the third video shows a concrete example on how to calculate the posterior, which supports in achieving the second learning goal: calculation.

Problem Solving

To reinforce Bayesian reasoning through hands-on experience, we adapted the "Find the Submarine" exercise from Bárcena et al. [1] into an interactive, web-based activity. As shown in Figure 3.1, each round of the game presents two grids side by side: on the left, the "Previous State" (which is blank during the first round), and on the right, the "Current State."

Figure 3.1: Game homepage, displaying previous and current probability distributions.

At the start of each turn, learners place a ship icon in one of the nine sectors and then estimate the probability that a hidden submarine (which may or may not appear) occupies that same sector in the next round. To aid their calculations and reduce extrensic cognitive load, students have access to an integrated calculator. Once they have computed their posterior probability, they click the "Search Sector" button and enter their estimate. Scoring is designed to reward accuracy: 1 point for an exact match, 0.2 points for a near miss, and 0 points otherwise. By iterating through multiple rounds and immediately observing how their probability estimates compare to the true posterior, students develop an intuitive sense of how likelihoods and priors combine to form updated beliefs.

To encourage further exploration beyond the structured game rounds, we provide a "Sandbox Mode" (see Figure 3.2). In this mode, learners can freely assign any arbitrary prior probabilities across the nine sectors and then place their ship to see the resulting updated probabilities. This open-ended environment invites students to experiment with different prior and likelihood combinations, supporting discovery learning, in which they may, through trial and error, uncover or approximate Bayes' Theorem before it is formally introduced.

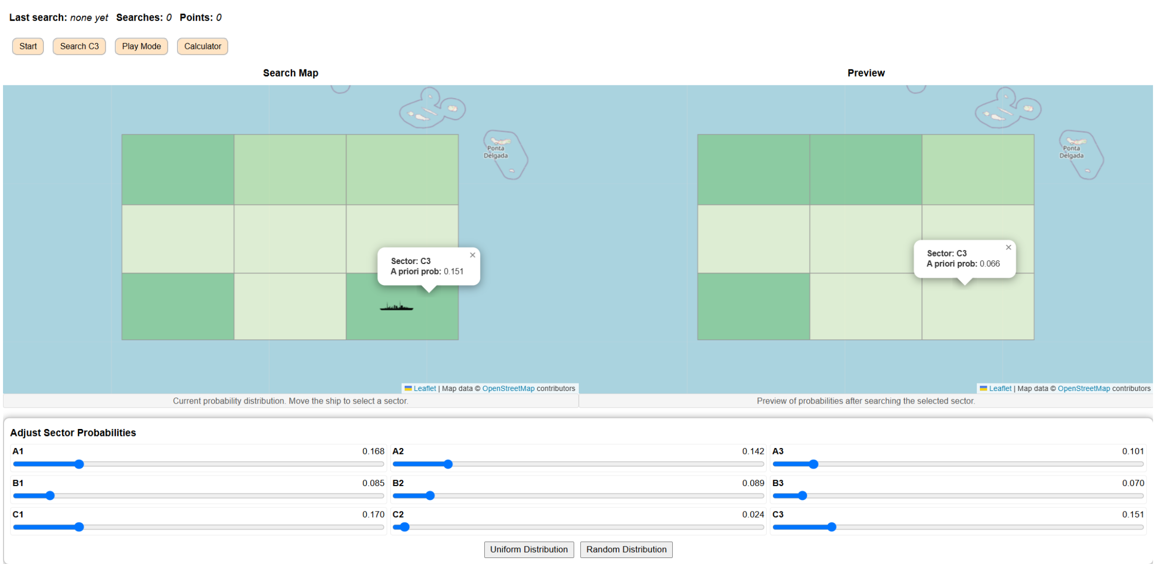

Figure 3.2: Sandbox mode, where students set arbitrary priors and observe the corresponding posteriors.

Importantly, this activity acts as a metaphor for the theorem itself: learners first make sense of the probability update through gameplay, and only afterward are introduced to the underlying mathematical structure. This aligns with constructivist learning theories and the principles of productive failure, where struggling with a problem before receiving instruction enhances deeper learning. Concretely, the problem solving activity focuses on computations that need to happen, hence supporting the second learning goal: calculation. Moreover, in the sandbox mode, participants need to make sense of the effect that changing a prior probability has on the posterior. Since, the example of searching a submarine is so realistic, participants are challenged to put meaning to the resulting posterior. E.g. they discover that it is not smart to search in a sector that has already been searched before (but without success). This fosters the third learning goal: interpretation.

Post-test

To complete the lesson, students take a post-test that mirrors the pre-test in difficulty and content areas while using different specific examples to avoid simple memorization. The post-test maintains the same structure (true/false and single-choice questions) and point distribution, allowing us to measure learning gains across each objective. Questions are designed to assess transfer of learning to novel situations, testing whether students can apply their understanding beyond the specific examples encountered during instruction.

Altogether, the lesson is designed to scaffold learning progressively through four main phases: pre-test, intuitive problem-solving activity exploration, guided instruction, and post-test. This structure can be implemented in two different sequences: PS-I (problem-solving before instruction) and I-PS (instruction before problem-solving). By comparing these two sequencing approaches, we aim to investigate how the order of these phases influences learning outcomes.

4 | Experimental Design

The primary research question we seek to answer is: *Which instructional sequence leads to greater learning gains when teaching Bayes' Theorem-Problem Solving before Instruction (PS-I) or Instruction before Problem Solving (I-PS)?*

To investigate this question, we adopt a between-subjects experimental design in which participants are randomly assigned to one of the two instructional sequences. Half of the participants are placed in the P-SI condition and the other half in the I-PS condition.

The independent variable in our study is the **instructional sequence**, with two levels: PS-I and I-PS. The dependent variable is the **learning gain**, which is defined as the difference in each participant's score between the pre-test and post-test. The tests are carefully designed to align with our four learning goals, covering recognition of applicable scenarios, calculation of posteriors, interpretation of results, and generation of real-world examples. The pre-test and post-test are designed at a similar difficulty level, each with a total score of 10 points. For each of the first three learning objectives, we designed one true/false question (1 point) and one single-choice question (2 points). For the final learning objective, we have included an open-ended reflection question. The full pre- and posttest can be found in Appendix A and Appendix B. In here, it can also be seen how every question of the pre- and post-test is linked to a learning objective.

To ensure validity, both groups receive identical instructional content and problem-solving tasks; only the order of activities differs. As the pre-test and post-test use parallel questions of equivalent difficulty, they are structured to allow direct comparison. This design enables us to isolate the effect of sequencing on learning outcomes while controlling that participants are exposed to the same content.

We will conduct learning gain comparisons (using statistical tests) to examine which sequence produces greater gains. Moreover, the moderation effect is tested, by looking whether the control variables age and gender have an influence on the learning gain.

5 | Implementation

To implement our lesson and experimental design, the Moodle platform was used. In here, the assessments (demographic survey, pre-test, post-test) were directly constructed. The demographic survey was by means of a 'feedback' assignment and the pre- and post-test by means of a Moodle 'quiz'. For the problem solving activity, a Moodle 'Assignment' was made in which the explanation of the activity was provided with a link to the game, hosted as web-application. Regarding the web-based game, it was primarily developed using HTML and CSS. The code was adapted from the open-source project BayesSim [5] by Fernando Tusell. The final version was deployed via GitHub, which provides free hosting for static web pages, allowing easy access and sharing of the game. Considering the instructional materials, a Moodle 'lesson' was created. Consisting of three steps. The first step was text that needed to be read. For the second and third step, Youtube videos were directly visible in Moodle. The time estimation of the whole experiment was as follows: 10 minutes for demographic & pre-test, 15 minutes for Problem Solving, 10 minutes for Instruction, 10 minutes for post-test. This distribution, ensured that the whole experiment roughly took 45 minutes, which made it easier to find participants.

To manage the groups, we created 20 unique Moodle accounts-10 assigned to the Problem Solving before Instruction (P-SI) condition and 10 to the Instruction before Problem Solving (I-PS) condition. Each account was pre-configured with the correct learning path according to the assigned condition. Participants were not aware of the sequencing difference, ensuring that expectations were consistent across groups.

Before releasing the experiment broadly, we conducted a pilot run with one test person to test the entire lesson flow, including the pre-test, instructional materials, problem-solving activity, and post-test. This pilot allowed us to confirm that the content delivery, transitions, and data collection mechanisms were functioning correctly. After verifying that the design worked as intended, we distributed the 20 accounts to our participants.

On May 2nd, we sent each participant their individual Moodle login credentials along with instructions and a link to the platform. Participants were given two full weeks to complete the lesson at their own pace. This time frame was chosen to accommodate different schedules and ensure participants had sufficient flexibility to engage thoughtfully with the material. All instructional elements-including embedded videos, interactive tasks, and assessments-were hosted and tracked within Moodle to maintain consistency and collect reliable data for later analysis. As the deadline to collect the experiment results approached, we noticed that a few participants had not completed the experiment on time. We did not want to pressure them, so we created additional accounts and invited two more participants to join the experiment. After May 16th, we began working on the data analysis. In class, we discussed our approach with Professor Patrick to ensure that we were on the right track.

Overall, the implementation reflects the intended lesson and experimental design, ensuring that the delivery was valid, consistent, and measurable across all participants.

6 | Participants, Data, and Analysis

6.1 | Data

The data that was collected from the participants consists of information regarding demographic background and regarding the learning gain. The learning gain information includes for both the pre-test and post-test, the total score on the test, the score per question of the test and the time it took to finish the test.

6.2 | Participants analysis

In total, 20 participants participated in our experiments, of which 10 in the I-PS and 10 in the PS-I part. As previously stated, all our participants fall in the category of having finished 'high school' and not having followed in-depth education on probability. Participants were found by asking people in the researchers' social networks that fall into the target group. Participants were given mock-up credentials in order to log into the Moodle platform to execute the experiment. In I-PS the gender distribution is 4 female, 4 male, and 2 unknown. In PS-I this is 5 female, 2 male and 3 unknown. Please note that 4 participants did not fill in the demographic survey, and hence there is no information about their demographic background. Moreover, one participant did not want to specify their gender. Additionally, in I-PS 5 students, 3 professionals and 2 unknowns participated whereas in PS-I 3 students, 5 professionals and 2 unknowns participated. The nationalities of our participants are Dutch(6), Chinese (5), unknown (4) Spanish(3), Luxembourgish(1) and American(1).

6.3 | Learning gain analysis

Exploratory Analysis

The learning gain is defined to be the difference between the total score of the post-test and the total score of the pre-test. It should be noted that no-one had a negative learning gain, i.e. everyone performed at least as good on the post-test as on the pre-test. The descriptive statistics of the learning gain can be found in Table 6.1. Additionally, a plot of the mean learning gain, including confidence intervals is shown in Figure 6.1. From this, intuitively, it can be seen that there is not a big difference in the learning gain between I-PS and PS-I, as their minimum, maximum and median are all equal. It can also be seen that the learning gain is more spread in the I-PS experiment than in PS-I. However, in order to formally state this, an ANOVA test is required.

condition	Min.	Q1	Median	Mean	Q3	Max.
IPS	0	0	2.5	2.6	5	6
PSI	0	1.25	2.5	2.5	3	6

Table 6.1: Descriptive statistics of learning gain

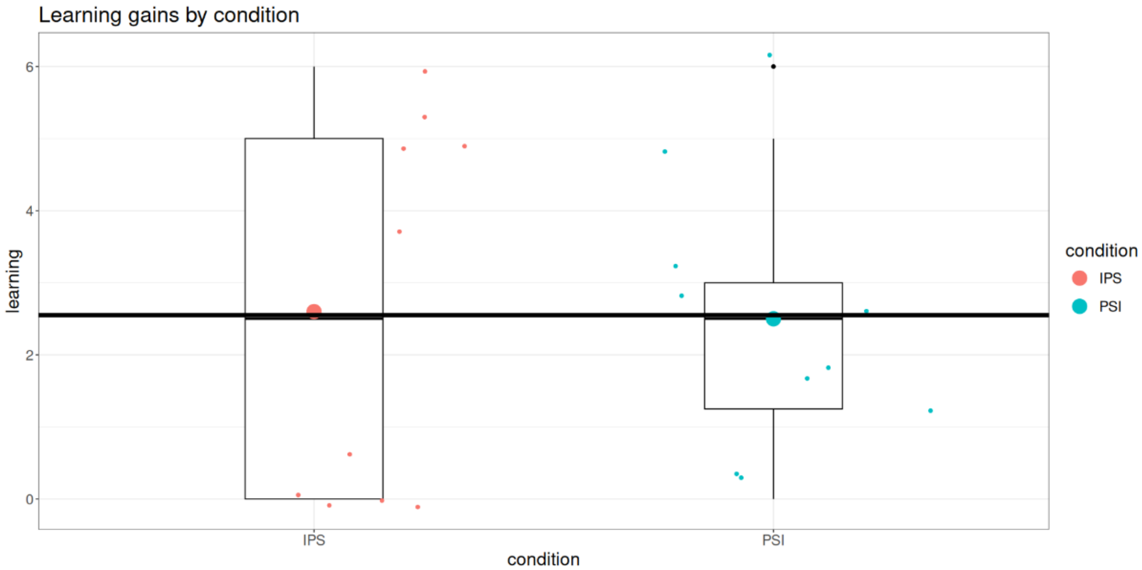

Figure 6.1: Average learning gain, given by condition

ANOVA & Kruskall-Wallis

An ANOVA analysis has been done where we looked at the learning gain based on the condition (order of instruction). This resulted in an F-value of 0.0095, with a p-value of 0.923. Clearly, we see that $p >> 0.05$, so we can not reject the Null hypothesis. Hence, we conclude that the results from the two experiments belong to two populations with the same mean. This means that the order of Problem Solving and Instruction does not have a significant effect on the learning gain.

It should be noted that ANOVA uses the assumption that the data follows a normal distribution and that there are roughly equal variances (i.e. homoscedasticity). By the Kolmogorov-Smirnov test, we found a p-value of 0.622 which is larger than 0.05 meaning we cannot reject the null-hypothesis. In this case, that means that it is likely that our residuals follow a normal distribution. However, the Shapiro-Wilks test gives us a p-value of 0.02 which is smaller than 0.05 meaning we reject the null hypothesis. So that

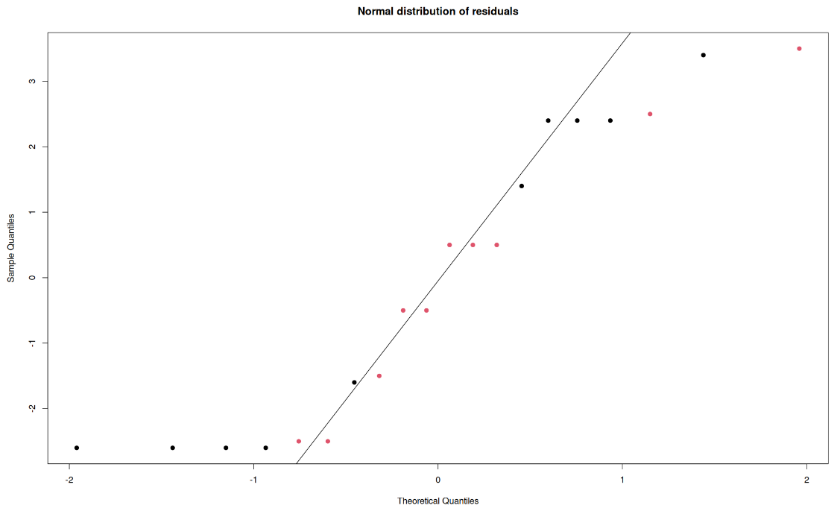

Figure 6.2: QQ plot residuals

means that our sample is not normally distributed. Moreover, from visual observations of the QQ-plot of the residuals (see Figure 6.2,) we can also notice that too many data points are off the center line, meaning our data is not normally distributed. Since the Shapiro-Wilk test is generally considered better when working with small sample sizes (which we are), we conclude that our data is not normally distributed.

Additionally, homoscedasticity has been tested by the Bartlett test. This gave a p-value of 0.4166, which is much larger than 0.05, meaning we can not reject the null-hypothesis. In this case, that means we can conclude that the variances are equal in both groups (PS-I vs I-PS).

Since our data is not normally distributed, the non-parametric equivalent of ANOVA: the Kruskall-Wallis rank test, has also been performed. This test gave us a p-value of 0.96 which is much bigger than 0.05, so we can not reject the null hypothesis. This means that the mean ranks of PS-I and I-PS are the same.

Moderation

Although we've shown that our independent variable (the order of problem solving) does not have a significant effect on our dependent variable (the learning gain), it could still be that other factors influence the learning gain. In other words, that moderation happens by our control variables. Our control variables are age group and gender. First, a check was done whether interaction happens between the independent variable and age, and between the independent variable and gender. For both control variables, there was no interaction. See Figure 6.3 to indeed notice that, when ignorning the 'none' category, there are no lines intersecting and hence no interaction. Since there is no interaction, type II (sums of squares) ANOVA tests have been executed. One with the independent variable and the control variable gender and one with the independent variable and the control variable age. There was no main effect of either gender or age on the learning gain, since the p-values were respectively 0.23 and 0.29. It should be noted that testing for moderation still used ANOVA tests. Fortunately, ANOVA is pretty robust to deviations from normality, meaning our moderation effect results are not too much affected.

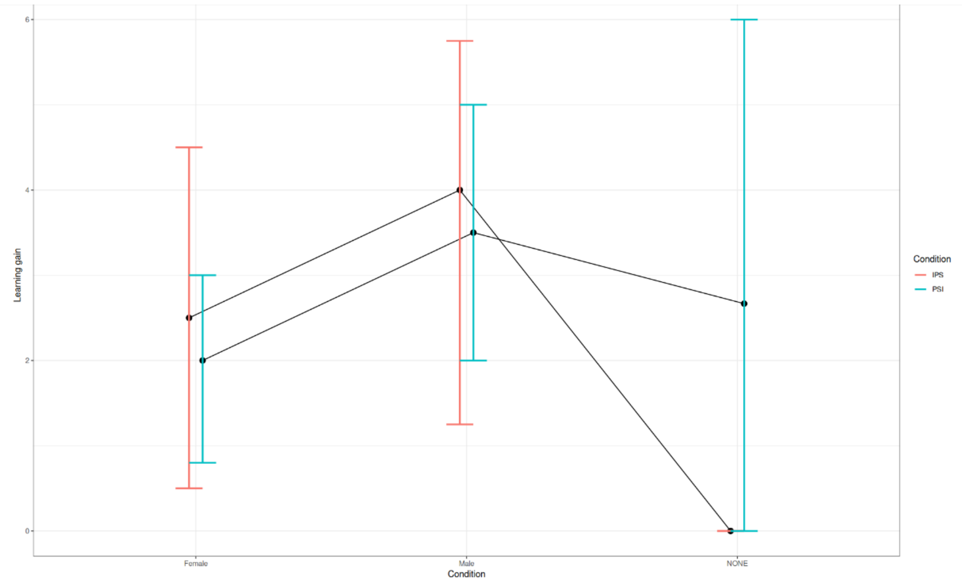

Figure 6.3: Learning gain categorised per gender

Power Analysis

Given a lack of effect and significant results, a post-hoc power analysis using G*Power [3] tool was conducted as can be seen in Figure 6.4. Specifically, we evaluated the power of our ANOVA with the following parameters: effect size f = 0.25 (medium-effect size, following Cohen's suggestions [4]), significance level $\alpha = 0.05$, total sample size = 20, and number of groups = 2. This resulted on a power estimate of $1 - \beta = 0.185$.

This result implies that, given our current sample size which is 20, we had only an 18.5% chance of detecting a medium-sized effect which is a very low percentage. A larger sample would be necessary to robustly evaluate the effect of sequencing on learning gain.

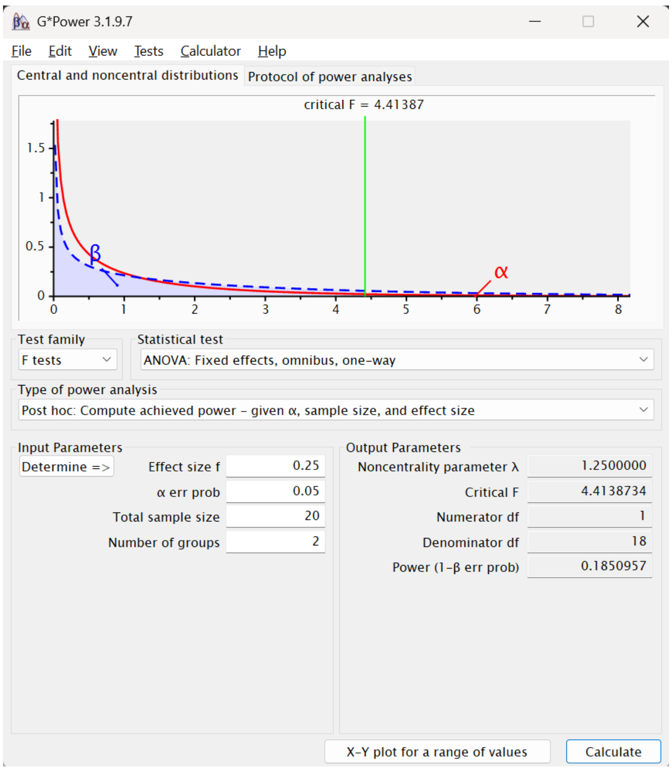

Figure 6.4: Power analysis using G*Power tool

7 | Conclusion and Discussion

Conclusion

The analysis of our data has shown that the order of problem solving and instruction does not have a significant effect on the learning gain when applied to the topic of understanding, and applying Bayes' theorem. In addition, both age and gender do not significantly influence the learning gain.

Discussion

However, due to our small sample size ($n = 20$) it cannot be claimed that 'Problem Solving before Instruction' will never have a positive impact on the learning gain. A bigger study, including a more diverse participant group should be held in order to be able to claim this. This is further supported by our post-hoc power analysis, which revealed a low power of only 18.5%, indicating a high risk of type II error. Additionally, no mediation analysis has been done as not enough information was available to facilitate this analysis. To allow future research to perform mediation analysis, the lesson design should focus on collecting information on e.g. the time spent on PS or the number of trials done during PS.

Moreover, it could be the case that our lesson design does not fully support productive failure. Our web-app in which the problem solving activity took place was not as intuitive to use as we hoped for as some of the participants mentioned that they struggled with using it. To overcome this issue, To address usability concerns, we could improve the UI/UX of the web app. One participant noted that it was difficult to find the start button; to remedy this, we can make the button more visible or enlarge its size. Alternatively, we could create a short video tutorial to guide users on how to navigate and use the web app effectively. Additionally, some of the participants were not fluent in English, which added more extrinstic cognitive load. To avoid cognitive overlad, we always want to minimize extrensic cognitive load. Hence, this was an undesired aspect of our experiment. Laslty, a reflection component could have been added to the problem solving activity, such as asking students to justify their choice or revisit previous decisions. By adding a reflection component, the game encourages metacognitive engagement, transforming experience into explicit understanding.

8 | Bibliography

[1] M. J. Bárcena, M. A. Garín, A. Martín, F. Tusell, and A. Unzueta and. A web simulator to assist in the teaching of bayes' theorem. *Journal of Statistics Education*, 27(2):68–78, 2019.

[2] Morris H. DeGroot and Mark J. Schervish. *Probability and Statistics*. Pearson Education, Boston, 4th edition, 2012. Discussion of Bayes' Theorem.

[3] Franz Faul, Edgar Erdfelder, Axel Buchner, and Albert-Georg Lang. Statistical power analyses using g*power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, 41:1149–1160, 2009.

[4] Stephanie Glen. Cohen's f statistic: Definition, formulas, 2025.

[5] Fernando Tusell. Bayessim: A web simulator to assist in the teaching of bayes' theorem. https://github.com/FernandoTusell/BayesSim, 2019.

A | Appendix A — Pre-Test Assessment Questions

Note: The pre-test and post-test are designed at a similar difficulty level, each with a total score of 10 points. For each of the first three learning objectives, we designed one true/false question (1 point) and one single-choice question (2 points). For the final learning objective, we have included an open-ended reflection question.

Learning Goal 1: Recognition

Question 1 (True/False):
Bayes' Theorem is appropriate for use only when the probability of new evidence is independent of the hypothesis being tested.
Answer: False

Question 2 (Single Choice):
Which of the following scenarios best illustrates the appropriate use of Bayes' Theorem to compute a posterior probability?

A. A person flips a fair coin 100 times and counts the number of heads.

B. A company estimates the average number of products sold per day over a year.

C. A weather app predicts rain tomorrow by combining prior weather patterns and new satellite data.

D. A student calculates the average of their last five test scores.

Answer: C

Learning Goal 2: Calculation

Question 3 (True/False):
A doctor knows that 1% of the population has a rare disease. A test for the disease is 99% accurate (true positive and true negative). If a person tests positive, the probability they actually have the disease is greater than 90%.
Answer: False
Explanation: Due to the low base rate, even with a 99% accurate test, the posterior probability is much lower (around 50%).

Question 4 (Single Choice):
A certain disease affects 1% of the population. A test correctly identifies 90% of those with the disease, but also gives a false positive result 5% of the time. A person is randomly selected and tests positive. What is the probability they actually have the disease?

A. 15%

B. 50%

C. 90%

D. 95%

Answer: A

Learning Goal 3: Interpretation

Question 5 (True/False):
If a test is 100% accurate, and someone tests positive, then they definitely have the disease.
Answer: False
Explanation: The prior probability matters too.

Question 6 (Single Choice):
A medical test is positive 90% of the time when a person has the disease and 10% of the time when they do not. The disease affects 1 in 100 people. Someone tests positive. Which of the following is closest to the chance they have the disease?

A. 90%

B. 50%

C. 10%

D. 1%

Answer: C

Learning Goal 4: Creation

Question 7 (Open-Ended Reflection):
Create a real-world example where you apply Bayes' Theorem to update a probability based on new evidence.
Sample Answer: A student applies for a scholarship. Historically, 10% of applicants are successful. After an interview, the interviewer gives a positive report. Data shows that 80% of awarded students had positive interviews, but only 20% of rejected students did. Bayes' Theorem can be used to update the student's chance of success based on this new evidence.

B | Appendix B — Post-Test Assessment Questions

Learning Goal 1: Recognition

Question 1 (True/False):
Bayes' Theorem can be used to calculate the probability of rain tomorrow based on today's weather forecast and historical weather data.
Answer: True

Question 2 (Single Choice):
Which of the following scenarios is **not** suitable for using Bayes' Theorem?

A. Estimating the chance a patient has a disease after a positive test.

B. Calculating the area under a curve using known values.

C. Determining the probability a student cheated on an exam based on suspicious behavior.

D. Updating the probability of finding a lost object after a failed search.

Answer: B

Learning Goal 2: Calculation

Question 3 (True/False):
A factory has two machines: A (30% production, 5% defect rate) and B (70% production, 10% defect rate). If an item is defective, the probability it came from Machine A is less than 50%.
Answer: True

Question 4 (Single Choice):
A rare disease affects 1 in 1,000 people. A test for the disease is 99% accurate. A randomly selected individual tests positive. What is the probability they actually have the disease?

A. 99%

B. 90%

C. 50%

D. 9%

Answer: D

Learning Goal 3: Interpretation

Question 5 (True/False):
In a criminal case, a computed posterior probability of 50% means the suspect is equally likely to be guilty or not guilty.
Answer: True

Question 6 (Single Choice):
In a criminal investigation, DNA evidence matches a suspect. The prior probability of guilt is 1%, and after updating using Bayes' Theorem, the posterior probability is 50%. What does this mean?

A. The suspect is definitely guilty.

B. There's now a 50% chance they are guilty.

C. The DNA evidence proves innocence.

D. The prior information was irrelevant.

Answer: B

Learning Goal 4: Creation

Question 7 (Open-Ended Reflection):
Create a real-world example where you apply Bayes' Theorem to update a probability based on new evidence.
Sample Answer: A student applies for a scholarship. Historically, 10% of applicants are successful. After an interview, the interviewer gives a positive report. Data shows that 80% of awarded students had positive interviews, but only 20% of rejected students did. Bayes' Theorem can be used to update the student's chance of success based on this new evidence.