Structured vs. Unstructured AI Prompting: Effects on Learning Outcomes and Efficiency in an Undergraduate Statistics Course

Sam Otto, Meghana Kolar. Immanuel Williams, Statistics.. California Polytechnic State University, San Luis Obispo

CAL POLY
Bailey College *of* Science & Mathematics

Introduction

AI is increasingly used by students as a study aid. The purpose of this study was to test how influential prompt engineering is on learning through a pre and post-test experimental design. Students in an undergraduate statistics course were assigned to use either unstructured prompting, RTF, or a control group with no AI. Their efficiency, perceptions of experience, and accuracy were studied.

Table 1: Pre-test accuracy by group. Accuracy = n_correct_pre / (n_correct_pre + n_incorrect_pre). Only consenting students included.

Group	n	Mean	SD	Median	Min	Max
No AI	15	0.900	0.134	0.917	0.500	1
AI — No Structure	16	0.906	0.085	0.917	0.667	1
Structured AI	15	0.956	0.062	1.000	0.833	1

At around 15 students per group, we had a total sample size of 46 students. They scored, on average, around 92% on the pre-test, and 75% on the post-test.

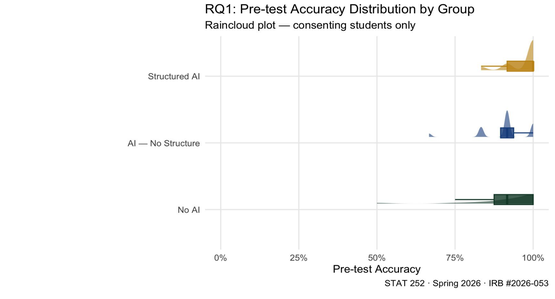
Table 2: Post-test accuracy by group. Accuracy = n_correct_post / (n_correct_post + n_incorrect_post). Only consenting students included.

Group	n	Mean	SD	Median	Min	Max
No AI	15	0.764	0.178	0.818	0.364	1
AI — No Structure	16	0.750	0.147	0.727	0.455	1
Structured AI	15	0.739	0.181	0.727	0.455	1

Methods

EXPERIMENTAL DESIGN:

- (1) Pre-test conducted to test base-level ANOVA knowledge (11 questions)
- (2) 20 minute teacher-led instruction in class to teach ANOVA
- (3) 20 minute study period. Study methods based on experimental group, and practice activity given to all students.
  - RTF group had prompts provided to plug into ChatGPT edu
  - Unstructured AI group had free reign
  - No AI group had to look over notes taken in class
- (4) Post-test (12 questions)
- (5) Post-activity survey
  - Feedback form, turning in Chat GPT history.



Results

Overall, we found no statistical significance between these groups that showed a difference in score based on study method. To further look at these results, we based our analysis on three research questions:

RQ 1

How does the distribution of pre-test performance vary across the three groups, and are there any baseline differences in student ability before the intervention?

RQ 2

How do post-test outcomes differ across the three groups after adjusting for baseline pre-test performance, and does the Structured AI group demonstrate significantly greater learning gains compared to the other groups?

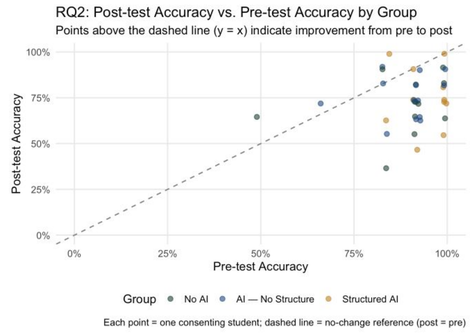
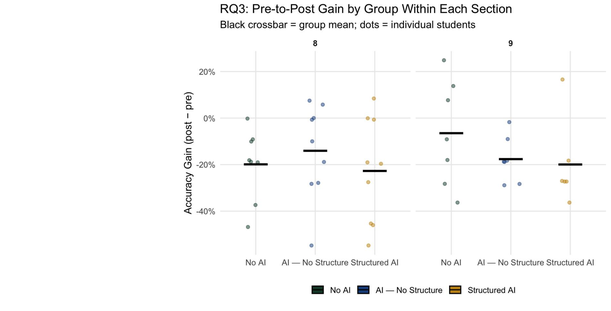


Table 6: Planned pairwise contrasts derived from ANCOVA EMMs. Contrast codes: No AI = -1, AI No Structure = 0 or ±1, Structured AI = +1. Positive estimate = Structured AI scored higher. Unadjusted p-values reported.

contrast	estimate	SE	df	t.ratio	p.value
Structured AI vs No AI	-0.0321	0.0639	42	-0.5026	0.6179
Structured AI vs AI No Structure	-0.0176	0.0626	42	-0.2810	0.7801
AI No Structure vs No AI	-0.0145	0.0612	42	-0.2372	0.8136

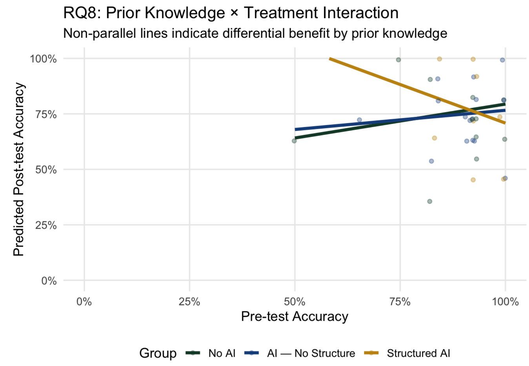
RQ 3

Within each section, do students in different groups show different rates of improvement in accuracy from pre to post, and which group has the largest improvement?



Future Directions

We could apply this ideology to a data science course or teach different subjects, such as joins and pivots, or exploring different packages in R studio. There are also a multitude of prompt engineering techniques we can use, such as tree of chain or JSON, since the graph below depicted an interaction between the Structured AI group and prior-knowledge. In addition, to increase the robustness of this study, we'd like to conduct this experiment again, strengthening our experimental design and perfecting our questionnaires.



Limitations

- The sample size of this experiment was relatively small, with around 15 participants in each group and 46 in total. This makes it hard to see true effect.
- Issues with the implementation of the experiment- keeping the instruction exactly the same for two sections was difficult.
- Controlling for how students used their study times- the exact prompts used, techniques used while using unstructured AI were hard to control for.
- Students were assigned randomly to conditions instead of based on pre-test score.

References

Ligot, D. V. (2026). From Role-Task-Format to SMART and Beyond: A Qualitative Taxonomy of Prompting Frameworks in Applied AI Practice. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.6130746