

PSYCH 10 Final Project · Fall 2026

Week 5 · M4 · Deciding on a research question

Full name (in CAPITAL LETTERS) _____ SUNet ID (the part before @stanford.edu) _____ Section (e.g., TUES 9:30) _____
Please write legibly so the teaching team can log your attendance: if we can't read your name, we can't give you credit.

Please hand this sheet in to your TA at the end of section (each student hands in their own copy). A copy is on the course website (psych10.github.io).

What you'll turn in

One member from each group will submit two files on Gradescope (the link is on Canvas): `M4-questions-and-dgps.Rmd` and the PDF it knits to. Again, remember that only ONE member submits—but that they must add the other members to the submission.

Steps

You'll do most of this in section with your group, and finish it together by Sunday. The `.Rmd` has a place for every question on this sheet.

1. **Set up your shared folder.** Your group should have received a link to your group's shared Google folder this week. (If not, please check in with your TA). The best way to use this folder is to:
 - download the folder and save it in the `psych10` or `stats60/` folder that you made a few weeks ago.
 - download `M4-questions-and-dgps.Rmd` from the course website and put it into the `milestones/` folder.
 - work on this `.Rmd` file in the folder, and decide ahead of time who in your group will be responsible for uploading the up-to-date version into the shared folder at the end of class.
2. **Commit to one claim and one dataset.** Start from the pairs you wrote down in M3 and your TA's comments. Discuss which claim and dataset you should choose, and then write down the claim you've chosen, the dataset, and a sentence on why your group chose this pair. From now on, this is your group's project.
3. **Transfer the code to access the chosen dataset.** Move the chosen dataset into your new group folder (in `data/raw/`), and, if you have the code from M3, copy your dataset loading code into this milestone. You can look at the dataset again with functions like `str()` and `dim()`.
4. **Pose your first question: how strong is the relationship?** Most claims are about how two things are related: "people who sleep more get better grades," "life expectancy has gone up since the 1950s." Turn your claim into a question about how strong the relationship is between two variables in your data. For example: *How much higher is `LifeExp` in 2007 than in 1952?* or *How much do `grades` go up with each extra hour of `sleep`?* Then write down:
 - your question, in one sentence, naming both variables as they appear in `str()`;
 - which variable is the **outcome** (the one you're trying to explain) and which is the **explanatory variable**, and whether each is quantitative or categorical;
 - the word equation: *outcome = explanatory variable + other stuff*;
 - what you expect: which way the relationship goes, and roughly how strong it is, in the units of the outcome.
5. **Think of the processes that could produce the relationship.** Suppose you find the relationship you expect. Write down at least three **different** data-generating processes (DGPs) that could have produced it. For each DGP, write one sentence on what you would see in the data if that DGP were the one at work. If two of your DGPs tell the same story in different words, replace one. Come up with your own list first; then ask an AI tool for more, and add any good ones it suggests, marked as the AI's.
6. **Look through your data for variables that could help.** Go back to `str()` and read the dataset's documentation. Which other variables could help you tell your DGPs apart? A third variable that could influence both of yours is the most common example: if a country's region affects both its wealth and its life expectancy, a column for region (`continent` in `gapminder`) matters. For each variable you find, write its name, what it measures, and which DGP it could help you check. If what you need isn't in your dataset, say what it is and whether another dataset might have it.
7. **Pose your second question, building on one alternative DGP.** Pick the alternative DGP from step 5 that your group finds most interesting. Write a second question about what other information could help account for the relationship. For example: *Is the relationship between `gdpPercap` and `LifeExp` still there when we look within each `continent`?* or *Is the difference in*

grades between students who sleep more and less partly explained by how many hours they work? Name every variable your question needs, and say where it comes from (this dataset, or another one).

8. **Show your TA your two questions** before you leave section, and write down what your TA suggested.
9. **Knit to PDF in a fresh session** (Session → Restart R, then Knit → Knit to PDF). One member submits the .Rmd and the PDF on Gradescope and adds the other members.
10. **Upload your files to the shared folder.** Now that you're done with this milestone, the member responsible for uploading (step 1) should use their computer to upload the dataset into the `data/raw/` folder in your group's Google Drive folder, and the M4 .Rmd into its `milestones/` folder. From now on, whenever your group works on the project, re-upload every file you changed to the shared folder, so that everyone works from the latest version.

Using AI tools on this milestone

This week, use AI tools to help you read the dataset's documentation and, once you have written your own DGPs (step 5), to find more candidate DGPs. Write your own list before you turn to an AI tool, so that you can tell your group's ideas from the tool's suggestions. Before you write down a variable name that you got from an AI tool, please confirm that you can see that name in the `str()` output for your dataset. Please choose the two questions yourselves, as a group.

How it's graded (2 points)

	Points
The data loads from your group folder, and the file knits; one claim and one dataset, with the reason	0.2
Question 1: one sentence naming two variables as in <code>str()</code> ; the outcome and explanatory variable, with their types; the word equation; a prediction with a direction and a rough size	0.5
At least three different DGPs, each with what you'd see in the data if it were true; the AI's suggestions marked	0.5
Other variables that could help tell the DGPs apart, each named as in <code>str()</code> with the DGP it could check (or what's missing and where it might be)	0.4
Question 2: builds on one alternative DGP and names every variable it needs; your TA's suggestions	0.4