Week 11 Practice - Putting it all together!

Elena Leib & Willa Voorhies

This week, you will be putting into action everything that you have learned so far this semester in order to explore a new data set from Qualtrics by preparing it for analysis and plotting it!

Before we begin, here is some background information about the data. *Please read it carefully!*

These data were collected as part of a study on abstract reasoning in early adolescence[1]. *The participants were 5th, 6th, and 7th grade students*, who participated remotely in their Zoom classrooms in May 2021. All tasks were hosted on Qualtrics, and each student received a unique link to participate via email at the beginning of the class period. *The study had two conditions:* an experimental condition that aimed to improve abstract reasoning and an active control condition that was intended to have no effect on abstract reasoning.

As the final task in the study, participants completed the WISC matrix reasoning task, a standardized 35-item task that measures abstract reasoning. This task consisted of 29 increasingly difficult trials (items 7-35) in which participants saw a visual puzzle with one piece missing (Figure 1). Participants were asked to "identify the missing piece" that goes with the puzzle by selecting its corresponding number. There were always 5 choices (1-5) and an "I don't know" option (0). If participants did not submit a response for the trial after 90 seconds, then Qualtrics automatically recorded 0 for the response and advanced the participant to the next trial. In order to prevent participants from becoming too frustrated, the task automatically ended if a participant got four trials incorrect in a row. Any combination of four trials selecting the incorrect response, selecting "I don't know", and/or taking longer than 90 seconds on a trial ended the task. Therefore, not all participants completed all 29 trials.

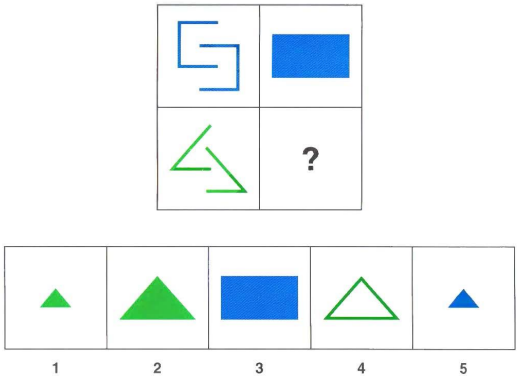

Figure 1. An example of a trial from the WISC matrix reasoning task. The correct answer is 2.

[1] These data were collected by Elena in May 2021. The study description and the data have been slightly modified for teaching purposes. Please do not share these data without written permission from Elena because they are unpublished and have been modified.

Your goal *is to examine whether matrix reasoning scores vary by grade (5-7) and/or condition (experimental vs. control).* For now, you will do this just by plotting the data. Although Qualtrics provides a total task score, you are not sure that it is accurate, so before you can look at the results, you will need to go back to the raw data, score the responses for each trial (as correct or incorrect), and calculate participants' matrix reasoning scores (i.e., the total number they got correct). Oh, and you're presenting these data in your lab meeting on Thursday, so chop chop!

Data

Your collaborator downloaded the data for this task from Qualtrics and sent it to you (`wisc_raw.csv`). To help you, your collaborator has also provided the response key for the matrix reasoning trials (`wisc_key.csv`) and a data codebook (`wisc_raw-codebook.xlsx`) to help you decode the column names in the data file.

Exercise Overview

Here is an outline of what you need to do in 4 parts. To help you, we have provided an example of what the data should look like at the end of each part.

On subsequent pages we provide scaffolded steps for accomplishing the goals below in each part, but we encourage you to try to figure out the steps on your own and use the scaffold as little as possible.

1. **Part 1:** *Read in and prepare the data*

   i) Tidy-up the dataframe
   - After you read in the data, look at it and ask yourself, "what do I need to do to clean up this data frame"?
   - If you're not sure you've cleaned everything up, `wisc_part1.csv` shows you what the data should look like after completing all the steps in part 1
   - Be sure to take a look at the codebook (`wisc_raw-codebook.xlsx`) to figure out what each of the columns are and how they should be processed

   ii) Familiarize yourself with the sample, the groups, and the conditions.

2. **Part 2:** *Score the data*

   i) Score the trial-level accuracy
   - The correct answers for each trial are given to you in `wisc_key.csv`
   - `wisc_part2-1.csv` - what the data should look like after scoring the trial-level responses.

   ii) Calculate total matrix reasoning score for each participant.
   - `wisc_part2-2.csv` - what the data should look like after calculating the matrix reasoning scores.
   - Remember, you aren't sure if Qualtrics scored your task correctly so compare your score to the score Qualtrics calculated.

3. **Part 3:** *Plot*
   i) Check to see if there are any outliers and remove them.
   ii) Did our experimental condition influence reasoning performance?
   - These plots should be lab meeting ready!

4. **Part 4**: *Test for true differences between conditions.*
   i) Create the same plot as above but now show an example of the relationship with matrix reasoning score if condition is assigned randomly.
   - This plot should be lab meeting ready and have the same layout and theme as the plots above.

   ii) Explore whether differences in reasoning score between the experimental and control groups are due to true differences or random chance.

Helpful tips:

1. For each part look at the data and think, *"What is my end goal and what are the steps I need to do to get there?"* Then write those steps out as comments before you start and/or use pseudo-code to write out the logic of your code or before you start.

2. Don't worry if you don't remember the syntax for a function. That comes with time and practice. Feel free to refer to google, your notes, or use the R help function.

3. If you get stuck, you can always peek at the scaffolded questions or ask us for hints! Remember, we are here to support you as you work your way through and answer any questions!

Best of luck! And may the duck be with you! 🦆

Scaffolded steps start on the next page.

Part 1: Read in and prepare the data

1) Orient yourself with the three files you have been given

2) Read in the data

3) View your matrix reasoning data, check the summary and structure. What do you notice about the data set? What is going on with the types for each column and why do you think this happened?

4) Remove the “trash” rows of the data frame. You could use indexing or any other function you want. Did this change anything about the data types?

5) Keep only participants who completed the task in the data set

6) Make sure no participant completed the task twice! If they did, there would be a 1 in Q_BallotBoxStuffing. Remove any participants that did the task twice.

7) Select only our columns of interest and rename them so they have meaningful names. Be sure that the participant information (pid, grade, cond, totalDuration, Q.score) appears as the first set of columns.

8) Change the levels of `condition` to have meaningful labels

9) If there is an issue with the data types of the columns, fix it to make it so the types are what you would expect them to be and make any categorical variables into factors (*Hint: you can do this column by column, but if you want to try something new, look into* `across()`)

10) How many students do we have in each grade? In each condition? In each grade and condition pair?

Part 2: Score the trial-level responses and calculate matrix reasoning scores

Now the data frame is in good shape and we are ready to score the trial-level responses. How do we go about doing that? Before you think about the coding part of it, think through the logical steps and what barriers (if any) there are to doing that with our data as we have them now. It may help to outline what it is that you will need to do.

First hint: Think about how you might easily do this in Excel! How can we do that in R? *For more hints, see scaffolded steps on the next page.*

Part 2 scaffolded steps:

1) What is the current shape of the data? (i.e., How many rows per participant? Are the trials in rows or in columns?)

2) The shape that I currently have my data in is not particularly conducive for scoring each trial. Why might that be? What would be a better shape for the data?

3) Reshape the data

4) Remove all the trials that participants never saw (i.e., where the response is NA). We want to keep only the trials that participants actually attempted.

5) Score the responses. *Hint*: It might be helpful to add in the response key to the data frame in order to line up the response and the answer for each trial (i.e., in each row).

6) Next, make a new data frame that has one row per participant with their WISC score as a column. Also compute the total number of trials that each participant saw. Be sure that your new data frame still has all of our other columns of interest (pid, grade, condition, totalDuration, and Q.score)

7) Was the score that Qualtrics had calculated for us correct?

Part 3: Plot!

Now that the data are formatted correctly we can visualize them. Our first step is to check the general distribution of our measures and then we can look at group effects.

1) Look at the distribution of matrix reasoning scores and task duration. Are there any participants who stand out? Be sure that you are using your calculated matrix reasoning score, not the Qualtrics score.

2) Look at the matrix reasoning scores in each grade. Do any participants stand out in their group?

3) Remove any participants who don't appear to be doing the task.

4) Plot the relationship between condition and matrix reasoning score. Look at the effects for each grade separately.

5) Re-label your axis and set the theme and color scheme of your plot so that it is ready for your lab meeting presentation.

6) Save this plot as a function so you can use the same theme and aesthetic for future plots.

Part 4: Test for true differences between conditions

First hint: How might we be able to test if condition is a meaningful grouping variable? Are there any sampling techniques that might help us?

For more hints, see the scaffolded steps on the next page.

Part 4 scaffolded steps:

1) First, let's calculate the true mean difference between our groups. (i.e., mean.exp - mean.control). Save this value as a new variable. You can collapse across grades to start but if you have time, you can try to do this for each grade and see what happens!

2) Now, create a new column in your dataframe (`condition_shuffled`) that randomly assigns participants to the experimental and control conditions.

3) Calculate the mean difference between the shuffled groups. How does it compare?

4) Now create the same plot as in part 3 but now use the shuffled condition data. How does it compare?

5) To get a better idea of the possible distribution of scores if condition is assigned randomly, we're going to repeat steps 2&3 multiple times. Use the code you created to shuffle data and calculate the mean difference, but now perform these steps 1000 times. Save the mean difference each time.

6) Plot a histogram of the possible mean differences and compare these to your true mean difference value.