

Information for RTs

Overview: Your tasks

■ Hypotheses to test

All Research Teams (RTs) are asked to submit a PAP and analysis code to test all four hypotheses presented below using the SOEP data. RTs shall consider these four research hypotheses that they are testing as causal hypotheses when developing their analysis:

H1. *Having daughters leads to stronger support for gender equality (as measured by attitudes).*

That is, denoting X as a measure of support for gender equality,
 $(X \text{ for respondents with daughters}) > (X \text{ for respondents without daughters}).$

H2. *Having daughters leads to more left-leaning political preferences.*

That is, denoting Y as a measure of left-leaning political preferences,
 $(Y \text{ for respondents with daughters}) > (Y \text{ for respondents without daughters}).$

H3. *Having daughters leads to more gender equality in the household (as measured by behaviors or outcomes).*

That is, denoting Z as a measure of gender equality in the household,
 $(Z \text{ for respondents with daughters}) > (Z \text{ for respondents without daughters}).$

H4. *Having daughters leads to a larger increase in the support of gender equality for men than for women (with gender equality measured by attitudes).*

That is, denoting X as a measure of support for gender equality,

$$\begin{pmatrix} X \text{ for male respondents with daughters} \\ - X \text{ for male respondents without daughters} \end{pmatrix} > \begin{pmatrix} X \text{ for female respondents with daughters} \\ - X \text{ for female respondents without daughters} \end{pmatrix}.$$

■ Pre-analysis plan

We would like you to write a pre-analysis plan detailing the exact way how you will test each of the four hypotheses. In particular, for each hypothesis, please consider the following:

- Variables: Which variables will you include? How will these variables be coded— will you use them the way they are or recode them, and if so, how?
- Inclusion/exclusion criteria: Are there any observations you will exclude? If yes, which? Which years will you include/what period will you study?
- Exact analysis: Please write down the exact specification of the regression model, including the names of the dependent and independent variables, and whether you use robust standard errors or something else.
- Interpretation of results: Please specify which statistic you will take into account for testing the hypothesis. How do you interpret whether your test result is in line with the hypothesis or not?

There are some constraints on the planned analysis: The analysis has to be conducted as a regression analysis (not necessarily OLS), and there should be one regression table per hypothesis. In this table, there should be regression coefficients, standard errors, values of the corresponding test statistic, and p -values, as well as any additional hypothesis test results based on the regression equation.

■ Analysis code

During the project, you will be asked to submit your analysis code (Stata do-files). When you do this, please submit separate analysis codes (do-files) for each of the four hypotheses.

Along with the data you will receive four example do-files, one for each hypothesis, which you can take as a starting point to write your code. The data and example do-files are arranged in a typical folder structure for an analysis package. If you open the example do-files by double-clicking or via “open-with” and then choose Stata, they will automatically set the working directory to wherever you saved them, and the example code should run without any further adaptations. The example do-files suggest a basic structure for organizing your code and provide commented sections, e.g., highlighting the hypothesis test.

Most importantly, your analysis for each hypothesis should result in (i) a coefficient estimate, (ii) standard error, (iii) family of the test statistics (z , t , F , χ^2), (iv) residual and model DFs (if applicable), (v) p -value, (vi) number of observations, (vii) number of participants, and (viii) interpretation of directionality of the effect, which are exported to a results CSV file (one per hypothesis). In the example do-files, you will find some information on how to extract these statistics and export them to a CSV file. You should use Stata version 14 or later.

■ Sample of the data

You will, at some point, get access to 5% of the data. Note that the sample is drawn at the household level based on the household composition at the survey entry. You will be randomized to one of two groups:

- Some of you will have access to the 5% already when writing the pre-analysis plan. We will ask those of you in this group to submit the analysis code and results for the 5% sample along with the pre-analysis plan. The analysis code will later be implemented by the project coordinators on the remaining 95% of the data.
- Some of you will instead write the pre-analysis plan without access to any data. After submitting the pre-analysis plan, those of you in this group will receive 5% of the data, and you will then be asked to implement your pre-analysis plan as closely as possible for the 5% of the data for the four hypotheses. The analysis code will later be implemented by the project coordinators for the remaining 95% of the data.

You should not change the analysis described in the pre-analysis plan unless the specified analysis cannot be executed in the data. The project coordinators will also test if the submitted code produces the results specified in the PAP. If this is not the case, we will contact you and interact with you until the code produces the results specified in the PAP.

Please keep track of the total number of hours you spend on preparing the pre-analysis plan, writing and implementing the analysis code, and working with the results, as we will ask you about later in the project.

■ Delivery and the use of the SOEP data and materials

You will receive a file package that includes the data and example do-files directly from SOEP data distributors at DIW, Berlin. Before receiving the data, you will be asked to digitally sign a standard data user form. Only after signing this form will you be eligible to receive the materials. Please refrain from using or acquiring further SOEP data for this project. This also includes SOEP data that you might have already gained access to previously for another study. Your analyses must only be based on the SOEP materials provided to you for the project.

■ Surveys

We will ask you to fill out three online surveys after having registered for the project. The first and second surveys will be about your beliefs about the hypotheses and your experience with the project, while the third survey will be on peer assessments (more below).

■ Peer assessments

Each RT will be asked to rate 10 other randomly assigned pre-analysis plans in terms of how easy they think it would be to implement the analyses in the pre-analysis plan and how reasonable they think the analysis path is. All pre-analysis plans will be anonymized for the peer assessment.

■ Timeline

April 27, 2025	Registration deadline
May 5, 2025	Participation notification
May 5, 2025 — May 18, 2025	Survey #1 / Data use agreement
May 26, 2025 — August 1, 2025	PAP and code writing
August 11, 2025 — September 07, 2025	Survey #2
September 08, 2025 — October 19, 2025	Peer assessment of PAPs
Begin of 2026 (estimated)	First draft of the paper

Getting started — Working with SOEP data for #ManyDaughters

For #ManyDaughters, you will use data based on the German Socio-Economic Panel's 2020 data release (Goebel et al., 2019; Siegers, Steinhauer and Schütt, 2022). The German Socio-Economic Panel involves data collected between 1984 and 2020. The dataset you will receive is limited to the years 2000 to 2020 and includes all individuals who completed the individual questionnaire in at least one of these years ($n = 90,357$).¹ Some but not all of these respondents appear in several waves over this period. Similarly, not all variables are collected in every wave.

You will work with data collected through the SOEP-Core. In particular, you will get access to the following datasets: *pl*, *hl*, *biol*, *biobirth*, *bioparen*, *pgen*, *hgen*, and *ppathl*.

■ Helpful resources

SOEPcompanion

- The [SOEPcompanion](#) is a user-friendly guide containing information on the topics, survey design, samples, and data structure of the SOEP.
- It also provides practical tips for working with SOEP data.

Finding information on variables

- Finding variables with [paneldata.org](#).
 - [paneldata.org](#) is a good source of information for all datasets and variables of the SOEP.
 - It provides comprehensive information on all variables. This includes value labels, frequency counts, and the survey years in which a variable was included. For some variables, there is also information on the corresponding survey question or on how the variable was generated.
 - To find information about a specific variable, enter the variable name in the search bar [here](#).
 - You can also search for keywords and find all variables that contain that word in their variable label.

Exploring datasets

- Exploring datasets with [paneldata.org](#).
 - You can find an overview of all datasets [here](#). The list includes the datasets you will receive.

¹In principle, the individual questionnaire is only completed by participants over the age of 18. However, in some survey years, there are rare cases of younger participants who have also completed this questionnaire.

- By selecting a dataset, you can browse through all the variables it contains.
- Exploring datasets with the [SOEPcompanion](#).
 - You can find detailed descriptions of each dataset [here](#).
- Note that in this project, you are limited to the datasets mentioned above.

Finding questionnaires

- Finding questionnaires with [paneldata.org](#).
 - You can find an overview of all questionnaires from all waves [here](#).
 - You can search for specific waves, questionnaire types (e.g., Household or Individual), or interview modes (e.g., Computer-Assisted Personal Interview (CAPI) or Paper-and-Pencil Interview (PAPI)).
 - Most questionnaires are available in English and German. For more recent waves, there are also questionnaires that indicate the name and label of the corresponding variable for each question.
- Finding questionnaires on the DIW website.
 - Alternatively, all available questionnaires can also be found [here](#).
 - Select “Survey Instruments” as “Kind of Documentation”. You can then select the desired year and type of questionnaire.

Finding codebooks

- A codebook is a document accompanying a dataset that provides detailed information about its structure, content, and variables.
- Finding codebooks with [paneldata.org](#).
 - Select the dataset, for which you want to find the codebook [here](#).
 - If available, the codebook for the most recent wave will be attached.
- Finding codebooks on the DIW website.
 - All available codebooks can be found [here](#).
 - Select “Data Documentations” as “Kind of Documentation”. Then, you can select the desired year.
- Note that there are no codebooks for the raw datasets *pl*, *hl*, and *biol*. As these datasets contain unmodified responses from the individual, household, and biography questionnaires, you can consult the respective questionnaires to gain insight into the content of the variables.

References

- Goebel, Jan, Markus M. Grabka, Stefan Liebig, Martin Kroh, David Richter, Carsten Schröder, and Jürgen Schupp.** 2019. "The German Socio-Economic Panel (SOEP)." *Jahrbücher für Nationalökonomie und Statistik*, 239(2): 345–360.
- Siegers, Rainer, Hans Walter Steinhauer, and Johannes Schütt.** 2022. "SOEP-Core v37- Documentation of sample sizes and panel attrition in the German Socio-Economic Panel (SOEP) (1984 until 2020)." SOEP Survey Papers No. 1106.