Pre-Analysis Plan:

Perceived sustainable minimum wages – a test of randomization and question scales across countries

Elisabeth Beckmann* Melanie Koch†

November 2022

*Oesterreichische Nationalbank, 1090 Vienna, Austria;
Email: elisabeth.beckmann@oenb.at
†Oesterreichische Nationalbank, 1090 Vienna, Austria;
Email: melanie.koch@oenb.at

1 Introduction

A survey experiment is a social science experiment that is embedded into a survey. For such an experiment, the survey sample is randomly divided into treatment groups which then receive a slightly different version of the questionnaire. This allows for treatment effect analysis of the bits of the survey that were varied. Using information provision experiments in surveys has gained increasing popularity in the last decade (e.g. Haaland et al., 2022). In such experiments treatment groups receive different pieces of information before they are asked to report beliefs or make certain, often financial, decisions. In this way information and salience effects can be tested. Moreover, randomization is often used to test for order effects of questions or answer options.

Being part of a team that runs an international survey – the OeNB Euro Survey – we plan to implement different survey experiments in future waves. The coverage of ten different countries makes the OeNB Euro Survey a very promising setting for survey experiments. It opens the potential for researchers to conduct cross-country experiments that are greatly harmonized with respect to field time, sample size and further questions asked. In the 2022 survey wave, we implement a survey experiment, the aim of which is threefold.

First, we test the feasibility of different randomization approaches across countries and whether falsification or errors in implementation regarding randomization might be a concern. Field work is prepared in close collaboration with us. However, we are never present during the field phase in any country. We receive the data only about one month after field work is completed, which is the end of November 2022. Thus, this year's test is used to establish best practices for implementing a randomization across each country's sample and to check whether faked or error-prone randomization is something we have to worry about.

Second, we examine data falsification in terms of individual responses. If there is evidence of data falsification, these effects how interviewers or survey institutes behave, would influence and/or blur treatment effects in respondents' behavior.

Third, we test a standard assumption in survey research regarding scale transformation. It is common in surveys when asking respondents to report wages to give them the choice which reporting unit to use, i.e. wage per hour, month or year. If these questions are used in research projects, numbers are then often set to the same scale; for example, by assuming that monthly wage = reported hourly wage $*$ official number of full-time working hours per month (e.g. Le Barbanchon et al., 2019). In our treatment, we vary on which scale respondents are supposed to report a specific wage. This experiment allows us to analyze if the assumption of, e.g., linear transformation based on full-time working hours, is valid.

This pre-analysis plan explains the randomization strategies and treatment in detail. It then outlines our hypotheses and how we plan to test for data falsification as well as treatment effects on reported wages by the respondents.

2 Research question

We address two main research questions:

1. Are there hints that the randomization procedure and/or reported values are faked?
2. Is it sensible to transform reported wages on one scale to another scale using simple assumptions on legal working hours?

To maximize the opportunity to detect potential falsification, we designed a treatment that satisfies the following criteria:

- Randomization should be over a question which respondents have to answer not over a piece of information that does not require data entry
- The question should be asked to all respondents, i.e. no filtered question
- Respondents should have to provide a continuous number, i.e. no question with only a few answer options like "yes-no"
- The question should be relatively easy to answer to have a low share of nonresponse, e.g. no question on economic expectations or knowledge question

The intention is that treatment groups will produce distinct distributions of numerical values that are not equal to each other and no obvious, easily calculated linear transformation of each other, if truly answered by the respondents.

We think that asking about wage or income figures fulfills the aforementioned criteria. We expect that it is not sensible to transform wages to the same scale ex post with a simple assumption (i) because genuine survey responses will be affected by known response behaviour effects such a rounding, which in turn will affect transformation to different scales and (ii) the average respondent will struggle with the mental arithmetic required to transform hourly wages to monthly wages. Given that many respondents are not part of the labor force (anymore), we do not ask about personal wages or reservations wages. Instead, we ask about the minimum wage respondents perceive as sustainable to make a living in their region. Asking about perceived minimum wages has the additional advantage that this question - in contrast to personal income - is not sensitive to answer.

To make sure that all country institutes are theoretically able to implement the randomization without large problems, we decided to have an equal sample split between two treatment groups, A and B. The questions posed to the respective treatment groups are:

Treatment group A:

Q154) In your personal opinion, what is the minimum amount people living in your town/village should approximately earn per hour after taxes?
_____ [Local currency/hour]
Don't know 88888

ONLY in Bulgaria, Romania and Serbia: If Q154=88888

Q156) In your personal opinion, what is the minimum amount people living in your town/village should approximately earn per year after taxes?
_____ [Local currency/year]
Don't know 88888

Treatment group B:

Q155) In your personal opinion, what is the minimum amount people living in your town/village should approximately earn per month after taxes?
_____ [Local currency/month]
Don't know 88888

Respondents in three countries might get a second question if they are assigned to treatment group A. This happens when they answer "don't know" to the first question. We made this adjustment because the survey institutes in the three countries told us that thinking in hourly wages is not common in their countries and they expect to receive a large share of nonresponse to this question. In case the shares will be truly large, we use the question on the yearly wage level as backup.

3 Implementation

3.1 OeNB Euro Survey

The OeNB Euro Survey is an annually conducted survey of individuals in ten different countries in Central, Eastern and Southeastern Europe. It mainly covers topics around euroization, financial situation of the respondents and their financial decisions. Sampling for the repeated cross-section is conducted on three stages. The target population are all persons aged over 18, residing and – in some countries – being citizens of that respective country. The countries are Albania (AL), Bosnia and Herzegovina (BA), Bulgaria

(BG), Croatia (HR), Czechia (CZ), Hungary (HU), North Macedonia (MK), Poland (PL), Romania (RO) and Serbia (RS). In total, at least 1,000 individuals per country are interviewed. In each country, a local survey institute administers the questionnaire, which is ex ante harmonized.

3.2 Field phase

The survey experiment is embedded in the OeNB Euro Survey wave 2022. The field phase is aligned across countries, with the first country going into field on 27th September and the last on 10th October. Depending on the country, the planned field phase duration is between two and six weeks. Neither we nor any other members of the OeNB Euro Survey team are present in the field in any country. Coordination and interviewer preparation is solely commissioned to external survey institutes. However, the team provides interviewer guidelines and discusses issues that could arise in the field with the survey institutes.

As in previous waves of the survey, the OeNB Euro Survey team will not receive any data before 30th November.

3.3 Randomization

Randomization of the treatment groups is administered separately across countries. It is stratified by interviewer to ensure that we can control for interviewer effects. In all countries except Czechia and Poland, all interviews are conducted on tablets and thus computer assisted. In Czechia and Poland, some fraction will be paper based. Last year the share amounted to 25% in Czechia and 38% in Poland. In all countries, interviewers are in charge of reading out each survey question and collecting the answers respondents provide.

The exact randomization mechanism is not unified across all countries. In countries with better internet connection and more advanced survey programming, CAPI-randomization will work in the background and without any actions necessary by the interviewer **(approach 1)**. These countries are Hungary and Romania as well as Czechia and Poland for CAPI-interviews. In the other seven countries, at the beginning of each interview interviewers have to type in the running number of the interview in a given sampling point **(approach 2)**. The question is phrased, e.g., "How many interviews have you already conducted in this sampling point"? or "In this sampling point, what is the number of this interview"? This running number will then be used for randomization.

Given that some interviewers conduct only a few interviews, each randomization mechanism sets a strict alternating order between group A and B treatments. This ensures that each interviewer will have a sufficient mix between group A and group B interviews.

Randomization approach 1 automatically switches between group A and B. Approach 2 will switch between group A and B depending on whether the interviewer types in an odd or even number (0 is counted as even for this purpose). The assignment of group A and B to odd or even can be different between interviewers and sampling points. For the PAPI-interviews, paper questionnaires are prepared with sequence numbers and stacked in such an order that these will be alternating as well **(approach 3)**.

There are still some further differences between the three approaches. While in approach 1 only completed interviews are counted into the randomization, for approaches 2 and 3 also interrupted interviews are counted. Interrupted interviews are counted in these approaches to ease up the procedure for interviewers as they are used to counting in interrupted interviews as well. Still, interrupted interviews are rare.

Moreover, the counter for randomization in approach 1 and 3 runs continuously through the field phase for each interviewer whereas in approach 2, the counter is reset for each sampling point the interviewer visits.

The different ways to randomize between countries is deliberately accepted to test how permissive each method is for intentional falsification or faulty implementation of the randomization procedure and to evaluate each method in terms of practicability.

4 Hypotheses

4.1 Randomization and data falsification

In terms of randomization and/or data falsification, we first look at overall results within a country. We employ three common indicators for data falsification that are suited to detect falsification in numerical, continuous values:

- The data do not obey Benford's law
- The share of rounding is low
- The variation between responses collected by a single interviewer is low

We apply these three indicators only on the questions related to the randomization treatment. We do not expect that the randomization is faked or that the data we receive are in fact not answered by the respondents. Thus, our initial hypothesis reads as follows:

H1: Looking at three general indicators for data falsification, we find clear evidence for faked data.

CESEE countries have exhibited substantial and persistent regional disparities with regard to income and unemployment Smętkowski (2013). Although there has been some convergence with poorer regions on average growing faster than more advanced regions, there has also been a hysteresis effect especially in the aftermath of crises. Given that respondents are asked to report the minimum wage they perceive to be sustainable in their region, we expect responses to vary across regions and correlate with regional disparities known from macroeconomic data – provided there is no falsification.

This gives us hypothesis 2a and 2b, where we complement the survey data with exogenous information on regional disparities:

H2a: The mean and median reported perceived minimum sustainable wages are not correlated with regional GDP.

H2b: The mean and median reported perceived minimum sustainable wages are not correlated with average stable night lights.

Data falsification might occur more likely in one treatment than in the other. This could happen because, for example, the randomization did not work and everyone received the same treatment. Thus, randomization assignment and data entries for one treatment have to be made up completely. Or because one of the treatment questions was harder to answer than the other. That one question is harder than the other was explicitly mentioned by three countries' survey institutes. In such cases, only some entries might be overwritten if falsification is present.

Therefore, we want to compare if the previously mentioned indicators differ between treatments for the three countries Bulgaria, Romania and Serbia. However, some of the indicators do not work for the comparison because they cannot be applied reasonably to two-digit numbers. Hence, we concentrate only on two previously introduced indicators, response variation by interviewer and response variation by region. Since we know that in all three countries, treatment group A is the more difficult question, we use a onesided test:

ONLY for Bulgaria, Romania and Serbia:
H3: Looking at the checks for data falsification, we find clear evidence that data are more likely faked in treatment group A than in treatment group B

It should be noted that hypotheses 1-3 will be tested for each country separately. Thus, each of hypotheses 1 and 2 in fact present ten independent hypotheses and hypothesis 3 presents three distinct hypotheses. We adjust all significance values for multiple

hypothesis testing using sharpened q-values from Anderson (2008).

4.2 Survey experiment: Transforming wages to different scales ex post

In the literature, scales for wages are often unified by that one is a linear transformation of the other based on legal working hours. We expect, however, that different response scales are in fact not perfect linear combinations of each other. Respondents might not have legal working hours of full-time employees in mind when giving wage estimations. They might think about their own working hours that could differ from the legal number or they use what they think should be legal working hours as benchmark. Moreover, low numeracy and rounding effects can play a crucial role. Calculating what a monthly wage means in hourly terms and the other way around, requires time and computational skills that some survey respondents do not have. Hence, we formulate hypothesis 4 as follows:

H4: Reported hourly wages are a strict linear transformation of reported monthly wages

We add two supplementary hypotheses for the three countries that potentially ask for yearly wages. However, since yearly wages are only asked if respondents did not provide an answer to monthly wages, we do not know ex ante if the yearly wage sample will be large enough.

H4a: Reported monthly wages are a strict linear transformation of reported yearly wages

H4b: Reported hourly wages are a strict linear transformation of reported yearly wages

Again, hypotheses 4-4b will be tested for each country separately. Still, in contrast to the issue of data falsification, we want to interpret the results on all countries combined as well. The implications differ depending on whether in most countries wages on different scales are no linear transformations of each other or whether in most countries they truly are.

5 Methodology

Let y_{ci1} be the variable that captures the answer to question 154, y_{ci2} the answer to question 155 and y_{ci3} the answer to question 156 for individual $i = 1, .., N_c$ in country $c \in \{AL, BA, BG, CZ, HR, HU, MK, PL, RO, RS\}$. All $y_{cij,j\in1,2,3}$ are bounded from below

by zero. *treat* is an indicator variable that equals 1 if only question 154 was asked, 2 if question 155 was asked and 3 if question 156 was asked. Nonresponse in the form of "don't know" answers is originally coded as 88888. For all further calculations, we treat these answers as missing values. Calculations for question 156 in the three countries applicable are only conducted if sample size is greater than 30.

5.1 Methodology for testing hypotheses 1-3

5.1.1 Hypothesis 1

1. We test if treatment data obey Benford's law following Schräpler (2011). However, we do not look at each interviewer individually but at all interviews within a country. We look at the first digits of all $y_{cij,j\in1,2,3}$ combined, but not at other continuous variables. The test statistic, for each country c, is the sum of the squared difference between the observed frequency of a digit being the leading digit (h_{c_d}) and the expected frequency according to Benford's law (h_{b_d}) divided by the expected frequency for all digits from 1-9:

$$\chi^2_c = \sum_{d=1}^{9} \frac{(h_{c_d} - h_{b_d})^2}{h_{b_d}}$$

This statistic follows a χ^2 distribution with 8 degrees of freedom. We will reject that treatment data are distributed according to Benford's law, if the (MHT adjusted) p-value for the realization is smaller than 5%.

2. For rounding, we use monthly personal income data that were elicited in the OeNB Euro Survey in the previous wave as benchmark. Virtually all of these numbers are multiples of five or even ten (more than 99.5%), meaning their last digit is a five or zero. We call this benchmark variable y_{ci}^{prev}. For each country, we compare the shares of multiples of five for all $y_{cij,j\in1,2,3}$ to the share in y_{ci}^{prev} using t-tests. For hourly wages, the share of responses that are multiples of five will be lower, as hourly wages frequently are two-digit numbers. For example, in studies on inflation expectations, where answers are often two-digit numbers, rounding still occurs frequently but is far from universal. For the t-test, we therefore exclude two digit numbers that could occur for hourly wages.[1]

3. For our variation by interviewer measure, we use the variability method similar to Schäfer et al. (2004). We calculate the squared distance between each $y_{cij,j\in1,2,3}$ from its sample mean within a country $\frac{\sum_{i=1}^{N_c} y_{cij}}{N_c} = \overline{y_{cj}}$. Then, we sum these squared distances for

[1] If there is a known minimum hourly wage, we expect that instead of rounding to multiples of five, responses will be heaped around the minimum hourly wage.

each question by interviewer. The final test statistic for each interviewer is then the sum of these two (or three) sums combined. For each interviewer, we compare this realization to a bootstrapped distribution of the test statistic. The distribution is obtained by re-sampling 95% of the sample 1,000 times. We will assume that the area under the density curve left from the realization is the probability that an interviewer has not faked the data. As above, we use 5% as cut-off point to reject falsification given this indicator.

Overall, we reject hypothesis 1 if at least two of the three indicators do not point in the direction of falsification.

5.1.2 Hypothesis 2

We believe that perceived sustainable wages vary with regional development and the standard of living within a region. Therefore, for hypothesis 2, we test

if there is a significant positive correlation between all $y_{cij,j\in1,2,3}$ and two variables that approximate regional development: the regional GDP and average stable night lights in a region.

If the Pearson correlation coefficients between each $y_{cij,j\in1,2,3}$ and regional GDP in 2022 and between each $y_{cij,j\in1,2,3}$ and regional average stable night lights in 2022 are significant, we reject hypothesis 2.

5.1.3 Hypothesis 3

For hypothesis 3, we look at the variance in answers per interviewer and the correlation between reported wages and GDP in a region. Concerning the variance, we calculate the test statistic described under point 3 in 5.1.1 separately for the treatments. Then, we do not use the bootstrapped distribution as point of comparison but test if the statistics for the treatments are significantly different from each other. Concerning the correlation, we test for each treatment separately if there is a correlation between given answers in a region and the GDP in the region. If the correlation is significant for both treatments, we see this as sign that the correlation indicator does not point to falsification.

If both indicators do not hint to falsification as defined above, we reject hypothesis 3.

5.2 Methodology for testing hypotheses 4-4b

We want to test if monthly wages are a linear transformation of hourly wages (and yearly wages).

1. We start with a visual inspection. We plot the kernel density estimate for each $\sum_{i=1}^{N_c} y_{cij}$. Then, we compare the shape of distributions between $\sum_{i=1}^{N_c} y_{ci1}$ and $\sum_{i=1}^{N_c} y_{ci2}$ (and $\sum_{i=1}^{N_c} y_{ci3}$, if applicable) for each c separately. If the shapes of the distributions do not look similar, we see this as first sign that they are no linear transformation of each other.

2. Next, we will transform hourly and yearly wages to monthly wages, using two simple assumptions:

a. We assume that the official working hours for a full-time employee are 40 per week, that one month has four weeks and one year 12 months. Then,

$$y_{cia} = \begin{cases} y_{ci1} * 40 * 4 & \text{if treat} = 1 \\ y_{ci2} & \text{if treat} = 2 \\ \frac{y_{ci3}}{12} & \text{if treat} = 3 \end{cases}$$

b. For each c and j, we calculate the average as given by the respondents wihtin a country $\frac{\sum_{i=1}^{N_c} y_{cij}}{N_c} = \overline{y_{cj}}$. Then,

$$y_{cib} = \begin{cases} y_{ci1} * \frac{\overline{y_{c2}}}{\overline{y_{c1}}} & \text{if treat} = 1 \\ y_{ci2} & \text{if treat} = 2 \\ y_{ci3} * \frac{\overline{y_{c3}}}{\overline{y_{c2}}} & \text{if treat} = 3 \end{cases}$$

3. Given that we then have all wages on the same measurement unit, we perform Kolmogorov-Smirnov tests for the equality of distributions of y_{cia} by $treat$ and y_{cib} by $treat$.

4. We perform t-tests for the equality of means of y_{cia} by $treat$ and y_{cib} by $treat$.

5. We run regressions separately for each country c, in which we control for regional fixed

effects (Z_i) and interviewer fixed effects (J_i):

$$y_{cia} = \alpha + treat_i'\beta + Z_i'\gamma + J_i'\delta + \epsilon_i, \forall c \in \{AL, BA, BG, CZ, HR, HU, MK, PL, RO, RS\} \quad (1)$$

$$y_{cib} = \alpha + treat_i'\beta + Z_i'\gamma + J_i'\delta + \epsilon_i, \forall c \in \{AL, BA, BG, CZ, HR, HU, MK, PL, RO, RS\} \quad (2)$$

In case we detect significant differences for the distributions and means within a country and in case treatment effects are significant in regressions (1) and (2), we interpret this as sign to reject hypothesis 4 (4a and 4b likewise, if applicable) for a specific country.

6 Power

The sample size is predetermined by the conditions of the contract between OeNB and the opinion poll institutes carrying out the survey. In each country, a sample of at least 1000 individuals aged 18 or older will be interviewed. Thus, each treatment group will have a minimum of 500 observations in each country. From previous waves of the OeNB Euro Survey, we have data on net monthly personal and household income. We calculate the minimum income per region for each country.[2] To calculate the standard deviation, we use the bottom income decile for each country. Note, that this likely overestimates the standard deviation for our survey experiment as the large regional disparities likely translate to the bottom income decile. Therefore, our power calculations indicate the lower bound. Table 1 shows the results. We show minimum regional income for the poorest two regions and the richest region.

Based on these previous survey results we conduct power calculations. We calculate that a sample of 1000 respondents in total, i.e., 500 per group, would allow us to detect differences in minimum income on the scale of the differences between the minimum income of regions 1 and 2 with a power of above 65% for all countries except Albania, Bosnia and Herzegovina, and Romania, assuming a type-I error rate of 5%. It would allow us to detect differences between region 1 and 3 with a power of 90% percent for all countries.

[2] For all countries, except North Macedonia, we use the equivalent of NUTS 2 regions. For North Macedonia, we use NUTS 3 regions.

Table 1: **OeNB Euro Survey: Minimum regional income**

Country	Region 1	Region 2	Region 3	Standard deviation of bottom 10 percent income per country
	minimum reported household income			
BG	100	155	200	73
HR	800	1000	3800	708
CZ	8000	8670	11400	1817
HU	14000	27000	71125	15043
PL	460	645	1600	404
RO	30	50	200	130
AL	4000	5000	24000	5671
BA	50	61	300	86
MK	700	1350	3694	2609
RS	1500	2500	7000	4644

References

Anderson, Michael L., 2008, "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, *103* (484), 1481–1495.

Haaland, Ingar, Christopher Roth, and Johannes Wohlfart, 2022, "Designing Information Provision Experiments." *Journal of Economic Literature*, *Forthcoming.*

Le Barbanchon, Thomas, Roland Rathelot, and Alexandra Roulet, 2019, "Unemployment insurance and reservation wages: Evidence from administrative data." *Journal of Public Economics*, *171*, 1–17.

Schäfer, Christin, Jörg-Peter Schräpler, Klaus-Robert Müller, and Gert G Wagner, "Automatic identification of faked and fraudulent interviews in surveys by two different methods." Technical Report, *DIW Discussion Papers*, No. 441, 2004.

Schräpler, Jörg-Peter, 2011, "Benford's law as an instrument for fraud detection in surveys using the data of the Socio-Economic Panel (SOEP)." *Jahrbücher für Nationalökonomie und Statistik*, *231* (5-6), 685–718.

Smętkowski, Maciej, 2013, "Regional Disparities in Central and Eastern European Countries: Trends, Drivers and Prospects." *Europe-Asia Studies*, *65* (8), 1529–1554.