Innovation Growth Lab
IGL

Preparatory Action for the Creation of a European Community of Practice on Policy Experimentation

Policymaker survey – Statistical analysis plan for experimental components

19 September 2025

1 Introduction

1.1 About the programme

The "Preparatory Action for the Creation of a European Community of Practice (CoP) on Policy Experimentation" is an ongoing initiative aimed at strengthening collaboration among policymakers across Europe to develop, test, and scale innovative policy approaches in research and innovation.

This initiative has the following objectives:

- Map key stakeholders to identify individuals and organisations active in policy experimentation
- Analyse existing policy experimentation approaches to identify best practices
- Assess the feasibility of establishing a structured platform for knowledge-sharing and collaboration

Through stakeholder engagement, surveys, and policy experiments, the initiative will generate evidence on the benefits, challenges, and enabling conditions for policy experimentation. The findings will contribute to policy recommendations and a governance framework for a potential CoP, ensuring that European policymakers have the necessary tools, knowledge, and networks to integrate experimental approaches into decision-making effectively.

The programme of preparatory actions is funded by the European Commission's Directorate-General for Research and Innovation and implemented by Technopolis Group in partnership with the Innovation Growth Lab (IGL) and Arctik. More information is available on the Commission's website.

1.2 About the current survey

The Innovation Growth Lab (IGL) and Technopolis Group will conduct an online survey of policymakers and policy practitioners across Europe, to assess their current level of understanding of policy experimentation, the drivers of and barriers to use of experimentation, and their views on the potential for a community of practice.

The survey incorporates four randomised experiments:

- A conjoint (discrete-choice) experiment designed to elicit respondents' willingness to engage in experimentation and which conditions are conducive to this.

- An A/B test to determine whether respondents' perceptions of their understanding of experimentation is affected by whether they are prompted with a particularly rigorous definition of what is meant by experimentation.
- An A/B test to examine whether respondents are more supportive of policy experimentation (and expect others to be more supportive) if they are first told of the results of a previous survey in which a majority of the general public in several European countries expressed generally supportive views about experimentation.
- An A/B test to examine whether respondents are more supportive of policy experimentation (and expect others to be more supportive) if the term "randomised experimentation" is used, rather than "randomised controlled trials (RCTs)".

2 Description of the dataset

2.1 Respondent selection

The survey will be open to all those working on research or innovation policy in government or other public-sector bodies in the European Union and the UK.

The survey will be promoted to policymakers and policy practitioners through IGL's and Technopolis Group's networks. Survey respondents will be self-selected, but the implementers will endeavour to ensure that there is representation from at least six countries, with approximate balance between northern, southern, eastern and western Europe. The target is to reach a total of 400 respondents.

2.2 Survey implementation

The survey will be carried out fully online, using the Medallia platform and an interface prepared by IGL staff. The survey will be prepared in English, but Google Gemini will be used to translate the survey into other major languages in the EU (including French, German, Italian, Spanish and other languages if resources allow). These translations will be quality-checked by an IGL or Technopolis staff member who is fluent in the target language.

The discrete-choice experiment will be implemented using Medallia's built-in package for conjoint experiments. Respondents will be presented with pairs of scenarios for a programme evaluation, each consisting of six attributes, shown in the table below. The content of each attribute in each scenario will be selected at random from the levels in the right-hand column of the table. Respondents will be asked to select which of the two scenarios they find preferable. They will be asked to make five such pair-wise comparisons between scenarios.

Attribute	Levels
Purpose of the evaluation	Assess the programme's impact Optimise the way the programme is delivered Deepen understanding of the need and potential solutions
Type of evaluation	Asking for feedback from users and stakeholders Monitoring of changes in outcomes and performance

Attribute	Levels
	indicators Participants are randomly assigned to receive different forms of support (Randomised controlled trial (RCT))
Cost of evaluation	€50 000 (5% of programme budget) €200 000 (20% of programme budget)
Set-up time available for evaluation	3 months 9 months
Evaluation timeline	Results needed in 6 months Results needed in 2 years
Technical support on evaluation	Available from external experts None available

The A/B tests will be implemented by generating three binary random variables within the Medallia interface when each respondent begins the survey. These random variables will determine whether the respondent is in the treatment or control arm for each of the three A/B tests.

2.3 Limitations

This survey is carried out with a self-selected sample. Those who respond to the survey are naturally more likely to be more interested in experimentation, more likely to be in networks connected to IGL and Technopolis, and with more time and willingness to respond to a survey. We will not be able to quantify the magnitude of these selection effects, so we should be cautious in generalising from the findings.

Many of the survey questions focus on hypothetical scenarios or activities. It is not known how accurately responses would predict respondents' actual behaviour.

3 Data cleaning

All data recorded before the launch date of the survey were collected during the piloting phase. Data from the pilot phase will not be included in the primary analysis, but may be included in the dataset when running robustness checks on the results. (Note that the pilot data was collected before this analysis plan was finalised, so this robustness analysis could not be considered as fully prespecified.)

It is possible that some respondents may come from countries outside the EU or UK. Data from any such respondents will not be included in the primary analysis, but will be included when carrying out checks on the robustness of the primary analysis.

The survey interface has been set up such that respondents are required to give responses to all the survey questions relevant to this analysis. For this reason, missing data is expected only for cases in which respondents stopped responding to the survey part-way through. In these cases, they will be included in the analyses for questions which they responded to, but not for subsequent questions that were missed. No imputation of missing data will be carried out.

Respondents will only be included in the analysis of the discrete-choice experiment if they complete all five pairwise comparisons.

4 Analysis of discrete-choice experiment

We seek to estimate the average marginal component effect (AMCE) for each of the six attributes tested in the discrete-choice experiment, as well as certain of the average marginal component interaction effects (AMCIEs) between the attributes.

Our analysis follows Hainmueller et al. (2014) and Schuessler and Freitag (2020). Using their terminology, the characteristics of our discrete-choice experiment are:

- Each respondent is presented with $K = 5$ tasks (pair-wise comparisons between scenarios), in each of which they choose between $J = 2$ profiles.
- Each profile consists of $L = 6$ attributes, each of which has either two or three levels.

The AMCE for each level l, β_l, will be estimated using regression models of the form:

$$Y_{ijk} = \alpha + \sum_{l=1}^{L-1} \beta_l A_{jkl} + \epsilon_{ijk} \qquad (1)$$

where for each individual i, task k and profile j, Y_{ijk} is the outcome variable, a binary variable equal to 1 if the profile was chosen and 0 if not, A_{jkl} is a binary variable equal to 1 if profile j in task k has level l for attribute A, and 0 if not, and ε_{ijk} is a random error term. (One level of each attribute is omitted, as the reference level.)

The AMCIE for the interaction between element l of attribute A and element m of attribute B, γ_{lm}, will be estimated using regression models of the form:

$$Y_{ijk} = \alpha + \sum_{l=1}^{L-1} \beta_l A_{jkl} + \sum_{m=1}^{M-1} \delta_m B_{jkm} + \sum_{l=1}^{L-1} \sum_{m=1}^{M-1} \gamma_{lm} (A_{jkl} \times B_{jkm}) + \epsilon_{ijk} \qquad (2)$$

where B_{jkm} is a binary variable equal to 1 if profile j in task k has level m for attribute B, and 0 if not.

Ordinary least squares regression will be used for estimating models 1 and 2, for ease of interpretation of the results. Probit regression will also be used as a check on the robustness of these results.

Schuessler and Freitag (2020) argue that it is not necessary to use clustered standard errors. However, as a check on the robustness of the results we will carry out the analysis with the standard errors clustered at the respondent level.

We define the primary and secondary analyses for this experiment as follows:

Analysis	AMCE(s)	AMCIEs	Number of comparisons
Primary	Preference for an RCT as the type of evaluation, as opposed to the other two options	Interactions between the choice of an RCT and each of the other five attributes (including the two levels for 'purpose of evaluation' – i.e. 6 comparisons)	7
Secondary	Each of the other 7 AMCEs	Each of the other 17 interactions between attributes	24

Adjustment for multiple comparisons will be carried out separately for (a) the primary analyses, and (b) the primary and secondary analyses considered together. This adjustment will be done using the method of Benjamini et al. (2006), allowing for a false discovery rate of 10%.

We will also undertake exploratory analysis. For example, we will examine whether the primary AMCE and AMCIEs differ between those who believe that innovation programmes are generally more or generally less impactful (as collected in question 4.5 of the survey).

5 Analysis of A/B tests

The hypotheses tested in the A/B tests are as follows:

Experimental manipulation	Hypothesis	Outcome measure
Vary whether respondents read a rigorous definition of experimentation before or after asking respondents to rate their level of understanding of experimentation	Respondents will rate their understanding of experimentation as lower when they are prompted with a particularly rigorous definition of what is meant by experimentation.	Self-assessed level of understanding of policy experimentation, measured on a Likert scale from 0 (not at all) to 3 (very well)
Vary whether respondents read a short summary of results of a previous survey in which a majority of the general public in several European countries expressed generally supportive views about experimentation before or after asking them about their own view and the views they expect of service users of the acceptability of experimentation	Respondents will express stronger support for randomised experimentation themselves if they are first told of the results of the previous survey.	Level of own support for randomised experimentation, on a Likert scale from 0 (strongly against) to 4 (strongly in favour)
	Respondents will expect those who are served or most closely affected by policies to be more in support of randomised experimentation	Expected level of support for randomised experimentation among those who are served or most closely affected by policies, on a Likert scale from 0 (strongly against) to 4

Experimental manipulation	Hypothesis	Outcome measure
	if they are first told of the results of the previous survey.	(strongly in favour)
Vary whether the terms “randomised experimentation” or “randomised controlled trials (RCTs)” are used when asking respondents about their own view and the views they expect of service users of the acceptability of experimentation	Respondents will be more positive about experimentation when the term “randomised experimentation” is used, rather than the term “randomised controlled trials (RCTs)”.	Level of own support for randomised experimentation, on a Likert scale from 0 (strongly against) to 4 (strongly in favour)
	Respondents will expect those who are served or most closely affected by policies to be more in support of randomised experimentation when the term “randomised experimentation” is used, rather than the term “randomised controlled trials (RCTs)”.	Expected level of support for randomised experimentation among those who are served or most closely affected by policies, on a Likert scale from 0 (strongly against) to 4 (strongly in favour)

Hypotheses will be tested using regression models of the following forms:

$$Y_i = \alpha + \beta T_i + \epsilon_i \qquad (3)$$

$$Y_i = \alpha + \beta T_i + \gamma X_i + \epsilon_i \qquad (4)$$

where, for each individual i, Y_i is the outcome variable, T_i is an indicator variable defined to be equal to 1 if respondent i is in the treatment arm(s) being tested and zero if the respondent is in the treatment arm(s) against which that treatment is being compared, X_i is a matrix of covariates, and ϵ_i is a random error term.

Ordinary least squares regression will be used for estimating these models.

Estimates from the models including covariates (equation 3) will be treated as the primary (preferred) estimates from the analysis.

Adjustment for multiple comparisons will be carried out separately for the estimates derived from each of the two sets of regression models (that is, those defined by equations 3 and 4). This adjustment will again be done using the method of Benjamini et al. (2006), allowing for a false discovery rate of 10%.

The following will be included in the matrix of covariates used in the models defined by equation 3:

- Indicators of the country the respondent works in
- Indicators of the geographical remit of the respondents’ institution

- Indicators of the aspect(s) of R&I policy the respondent works on (funding, design/development, implementation, evaluation/analysis and/or administrative or technical support)
- Indicators of the policy or programme area the respondent works on.

In the second and third A/B tests, the respondent's trial arm in the other trial will also be added as a covariate. Due to limitations of statistical power, tests for the interaction between the two A/B tests will not be carried out.

As a robustness check, the analysis will also be repeated with the sample restricted to those who completed the survey. For this analysis, the following additional covariates will be added, for which data is collected in the penultimate section of the survey:

- Indicators of the respondent's job role
- Indicators of how long the respondent has worked in public administration
- Indicators of the respondent's education level
- Indicators of the respondent's age
- Indicators of the respondent's gender.

6 Statistical power

6.1 Discrete-choice experiment

Assuming that the survey reaches its target sample size of 400 respondents and each carries out five pairwise comparisons in the discrete-choice experiment, the effective sample size (as defined by Schuessler and Freitag) will be 4000. This will provide 80% power to detect:[1]

- AMCE of 5.5 percentage points for attributes with three levels (including our primary AMCE analysis), and 4.5 percentage points for attributes with two levels.
- AMCIE of 10.7 percentage points for the interaction between an attribute with three levels (including whether the RCT is selected) and one with two levels, or 13.1 percentage points for the interaction between two attributes each with three levels.

This analysis does not take account of the potential explanatory power of covariates – this would reduce the detectable effect size somewhat. However, it also does not take account of the correction for multiple hypothesis testing, which would increase the detectable effect size.

6.2 A/B tests

Again assuming that the survey reaches its target sample size of 400 respondents, the A/B tests will have 80% power to detect an effect of 0.28 standardised deviations of each of the outcome measures.

[1] Analysis carried out using the cjpowR package in R (Freitag & Schuessler, 2020).

Outcome measure	**Scale**	**Standard deviation in survey pilot data**	**Hence detectable effect size**
Self-assessed level of understanding of policy experimentation	0 (not at all) to 3 (very well)	1.16 points	0.32 points
Level of own support for randomised experimentation	0 (strongly against) to 4 (strongly in favour)	0.65 points	0.18 points
Expected level of support for randomised experimentation among those who are served or most closely affected by policies	0 (strongly against) to 4 (strongly in favour)	1.17 points	0.33 points

Again, this analysis does not account for the explanatory power of covariates, nor for the correction for multiple hypothesis testing.

References

Benjamini, Y., Krieger, A.M., & Yekutieli, D. (2006). Adaptive linear step-up procedures that control the false discovery rate. *Biometrika*. 93(3), 491–507. https://doi.org/10.1093/biomet/93.3.491

Freitag, M., & Schuessler, J. (2020). *cjpowR – A priori power analyses for conjoint experiments*. R Package. https://m-freitag.github.io/cjpowR_shiny/

Hainmueller, J., Hopkins, D. J., & Yamamoto, T. (2014). Causal inference in conjoint analysis: Understanding multidimensional choices via stated preference experiments. *Political Analysis*. 22(1), 1–30. https://doi.org/10.1093/pan/mpt024

Schuessler, J., & Freitag, M. (2020). *Power analysis for conjoint experiments*. https://doi.org/10.31235/osf.io/9yuhp