Supplementary Information/Online Appendix

Public Opinion on Welfare State Recalibration in Times of Austerity: Evidence from Survey Experiments

December 2021

Table of Contents

A Operationalization and summary statistics of key variables

Table A.1: Operationalization of variables used as covariates

Variable	Question Wording	Measurement
Age	What year were you born?	Numerical variable: 18-88
Female	What is your sex?	Binary: 1 = female; 0 = men or no answer
Married	What is your marital status?	Binary: 1 = married; 0 = not married
Children	How many children are living in your household under the age of 10?	Binary: 1 = children; = no children
Education	What is the highest level of education that you have achieved so far?	Categorical: 1 = primary / lower secondary; 2 = secondary; 3 = tertiary
Class*	What is/was your occupation?	Categorical: 1 = employer; 2 = middle class; 3 = working class; 4 = routine worker
Income	Please indicate the answer that includes your total household income in the previous month, after taxes and compulsory deductions?	Numerical variable by percentile: 1-10 OR categorical variable by percentile: 1 = 1st - 3rd; 2 = 4th - 7th; 3 = 8th - 10th
Retired	Which of these descriptions applies to what you have been doing for the last months?	Binary variable: 1 = retired; 0 = otherwise
Work contract	Based on three questions: (i) What is/was your occupation? (ii) What is/was the nature of your employment? (iii) What describes best your current working status?	Categorical: 1 = outsiders (part-time workers with < 30 hours, temporary workers, or unemployed); 2 = insiders (full-time permanent working contract); 3 = upscales (higher skilled professionals, large employers & business owners, and self-employed citizens); 4 = out of work (retired, homemakers, students)
Union	Are you or have you ever been a member of a trade union or similar organization?	Binary: 1 = current union member; 0 = otherwise
Partisanship*	Which party did you vote for in the last (COUNTRY) election in (MONTH/YEAR)?	Categorical: 1 = far left; 2 = center left; 3 = center right; 4 = far right; 5 = other party; 0 = abstention
Left-right scale	In politics people sometimes talk of "left" and "right". Using the scale below, where would you place yourself, where 0 means the left and 10 means the right?	Categorical: 0-4 = left; 5-6 = center; 7-10 = right

* = Please see below for further information on the exact operationalization.

Class was coded according to the following occupational groups. (1) *Small and large employers*: self-employed farmers, fishermen, professionals, and shop owners/ business proprietor. (2) *Middle class*: employed professionals, general management, middle management, employed position desk with upper secondary education, employed position travelling with upper secondary education, and employed position service with upper secondary. (3) *Working class*: skilled manual workers, supervisors, and unskilled workers with an upper secondary education. (4) *Routine workers*: unskilled worker without upper secondary, employed position desk without upper secondary education, employed position travelling without upper secondary education, and employed position service without upper secondary

Partisanship was coded according to the ParlGov database, which classifies parties into families by their position in an economic (state/market) and a cultural (liberty/authority) left/right dimension. The party families were then simplified into five political groups, as shown below:

Table A.2: Classification of political parties into five groups

	Germany	**Italy**	**United Kingdom**
Far right	Alternative für Deutschland	Lega Nord, Fratelli d'Italia	UK Independence Party
Center right	Christlich Demokratische Union, Christlich-Soziale Union, Freie Demokratische Partei	Scelta Civica, Nuovo Centrodestra, Forza Italia	Conservative, Liberal Democrats
Center left	Sozialdemokratische Partei Deutschlands, Grüne/Bündnis90	Partito Democratico, Radicali Italiani, Articolo 1/MDP	Labour, Scottish National Party, Greens
Far left	Die Linke	Sinistra Italiana, Movimento 5 Stelle, Rifondazione Comunista	
Others	Piratenpartei	Südtiroler Volkspartei	Plaid Cymru

Table A.3: Summary Statistics

Variable	N	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
Education spending (pooled)	3,600	6.309	2.602	0	5	8	10
Childcare spending (pooled	3,600	5.530	2.725	0	4	7	10
ALMP spending (pooled)	3,600	5.418	2.715	0	4	7	10
Split	3,600	2.498	1.117	1	1.8	3	4
Age	3,600	46.883	16.029	18	34	60	87
Female	3,597	0.513	0.500	0.000	0.000	1.000	1.000
Married	3,600	0.459	0.498	0	0	1	1
children_u10	3,600	0.209	0.407	0	0	0	1
Education_c	3,585	2.213	0.676	1.000	2.000	3.000	3.000
Class	2,999	2.186	0.826	1.000	2.000	3.000	4.000
Income	3,427	4.967	2.803	1.000	2.000	7.000	10.000
Retired	3,600	0.207	0.405	0	0	0	1
Union	3,546	0.272	0.445	0.000	0.000	1.000	1.000
Partisanship	3,335	2.163	1.302	0.000	1.000	3.000	5.000
Left-right scale	3,600	1.845	0.737	1	1	2	3
Work scale	3,289	2.534	1.067	1.000	2.000	3.000	4.000

B Case selection, sampling, and quality tests

B.1 Case selection

Table A.4: Case Selection

	DE	IT	UK
Welfare regime	Continental	Southern	Liberal
Varieties of capitalism	CME	MME	LME
Unemployment rate, 2017	3.7	11.2	4.3
Youth unemployment rate, 2017	6.8	34.8	12.1
Government debt (% of GDP), 2017	72	153	117
Education spending (% of GDP), 2016	3.6	3.3	4.2
Pension spending (% of GDP), 2016	10.1	16.2	6.2

B.2 Sampling

Our survey was fielded in three large European countries in January 2018: Germany, Italy, and the United Kingdom. In each country, 1,200 respondents were recruited to participate in the survey. A large online panel provided by *Qualtrics* was used with tens of thousands of panelists across all age brackets. Qualtrics only uses double-opt in online panels that allow their panelists to take a survey no more than once every two weeks. Respondents were then drawn from a pool of eligible voters in each country and the sample was representative of all eligible voters based on gender and age, meaning we introduced quota sampling based on age and gender.

Given our subgroups comparisons, we are particularly interested in having roughly representative income and partisan groups. Even without quotas sampling by income and ideology, our sample is already fairly representative in terms of both variables. Figure A.1 plots the share of respondents by income decile for all countries (left) and by country (right). Overall, our income variable is almost equally distributed across income decile, only with a slightly higher representation of low-income respondents compared to high-income respondents. While the German sample is quite well balanced across income decile, the Italian and British sample is slightly more skewed towards low-income respondents. However, the differences are overall not particularly large.

In addition, we compared the party vote share in the closest election to the share of voters in our survey (Table A.5). We distinguish between the far right, center right, center left, and the far left, as explained in more detail in Table A.2. Especially in Italy, our sample matches the actual vote share that these four blocks received very well. In Germany, far left and center left voters are slightly over-represented, while center right voters are underrepresented. In the United Kingdom, center right voters are also underrepresented while center left and far right voters are slightly over-represented. Overall, the survey vote share matches fairly well the actual vote share.

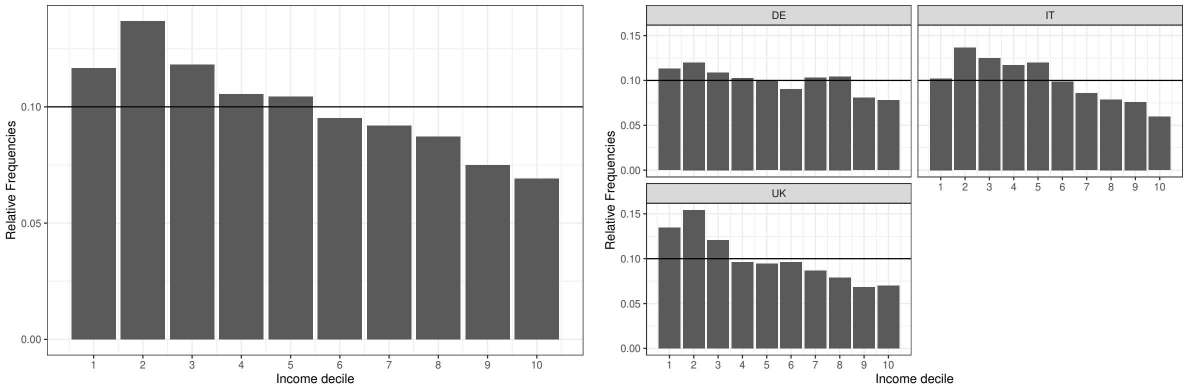

Figure A.1: Share of respondents by income decile (left: all countries; right: by country)

Table A.5: Survey vote share versus actual vote share

		Far Left	**Center Left**	**Center Right**	**Far Right**
	Election	9.2	29.4	43.7	12.6
DE	Survey	15.9	32.1	33.2	11.2
	Diff.	6.7	2.7	-10.5	-1.4
	Election	34.4	22.2	18.8	21.8
IT	Survey	38.5	22.1	15.6	19.9
	Diff.	4.1	-0.1	-3.2	-1.9
	Election	0.0	47.0	49.9	1.8
UK	Survey	0.0	51.6	40.9	5.7
	Diff.	0.0	4.6	-9.0	3.9

We also used weights to match the demographic characteristics of the country's population as closely as possible using entropy balancing (Hainmueller 2012). Please see Appendix F for a detailed explanation and the results using entropy balancing.

B.3 Quality checks and screening out of respondents

To ensure the quality of our sample, the survey included an attention check and a speeding check. We implemented a common attention check in the middle of the survey. Respondents were shown the following question:

We are interested in learning about your preferences on a variety of topics, including colours. To demonstrate that you have read this question, please go ahead and select both red and green from the alternatives below, no matter what your favourite color is. Yes, ignore the question below and select both of those options. What is your favourite color?

The answer options were: blue, red, purple, orange, yellow, green, brown, gray. Those who did not select the two correct color, were excluded from the survey. Additionally, we implemented a speeding check for the whole survey and one specifically for the conjoint experiment. Respondents who had a response time below 1/3 of the median response time were excluded.

C Instructions for the Conjoint Experiment

The full instructions for the conjoint tasks are shown below. First, respondents were presented with the following introduction to the experiment:

> **Please take your time and read the information below very carefully. It contains the instructions for the next part of the survey.**
>
> Every year the [COUNTRY] government spends money in a variety of different areas. We are interested in how you would like the government to change its spending pattern.
>
> We will now show you several proposals for possible changes to government spending in different areas. We will always show you two possible proposals in comparison. For each comparison we would like to know which of the two proposals you prefer. You may like both proposals or neither. In any, case please choose the proposal that you like the most. In total, we will show you five comparisons.
>
> The possible proposals only include changes with regard to a few selected types of government spending. Please assume that spending in all other areas is held constant. Please also assume that taxation and the level of government debt are held constant.
>
> People have different opinions about this issue and there are no right or wrong answers. Please always take your time when reading the proposals.

After this, the screen presented two reform proposals, as shown in Figure A.2. Respondents were asked five times to choose (i) between two packages ("choice variable") and (ii) to indicate how likely they are to support each of the proposals ("ranking variable").

Please carefully review the options detailed below, then please answer the questions.

Which of these proposals do you prefer?

	Proposal 1	**Proposal 2**
Child benefits	Decrease spending	No change
Old-age pensions	No change	Decrease spending
Education	No change	Increase spending
Training for the unemployed	Increase spending	No change
Unemployment benefits	Increase spending	Decrease spending
Childcare services	Decrease spending	Increase spending

Proposal 1

Proposal 2

How would you rate proposal 1 on a scale from 0 to 10, where 0 indicates that the government should definitely not adopt the proposal and 10 indicates that the government should definitely adopt it?

0 - Definitely not adopt	1	2	3	4	5	6	7	8	9	10 - Definitely adopt

How would you rate proposal 2?

0 - Definitely not adopt	1	2	3	4	5	6	7	8	9	10 - Definitely adopt

Figure A.2: Screenshot of a conjoint task presented to respondents

D Explanation of ridge regression to analyze the conjoint experiment

We propose ridge regression as a novel method to analyse conjoint survey experiments with dependencies. In our design, the values that each attribute can take are linearly dependent on the other attributes to ensure that the budget is balanced. Therefore, there is explicit confounding between the features of the fiscal packages, due to the restrictions to the randomization protocol that we introduced. We included these restrictions to increase external validity, as recommended by Hainmueller et al. (2014, p. 26). They argued that "researchers should employ restrictions and exclude attribute combinations whenever they are deemed so unrealistic that a counterfactual would essentially be meaningless." We took this recommendation to the extreme because the budget constraint of the public budget is binding: changing the government's expenditure or revenue on one dimension necessarily has consequences for another dimension. This poses difficulties for traditional regression analysis, but it ensures external validity.

To see this, imagine the alternative: We could have introduced combinations that are unrealistic to break the super-collinearity that exists in reality. This would have allowed us isolate the effect of each attribute independently of other attributes, but it poses a threat to external validity. Respondents would have probably most preferred atypical profiles, which increase government spending, decrease taxation, and cut government debt all at the same time. Given our interest in citizen's priorities in the face of hard budgetary tade-offs, this would not have been satisfactory. As Bansak et al. (2020, p. 24-25) argue, "the AMCE averages the effect of an attribute over two different distributions: the randomization distribution of the other attributes and the distribution of respondents." The AMCEs of a fully randomized design, therefore, are not of interest to us. We only wanted to make inferences about realistic budgetary combinations, i.e., those that are fully balanced.

The resulting experimental design necessitates a modelling approach which accounts for the design-based super collinearity. One common solution to enable OLS standard regression analysis in instances of super collinearity is to drop one of the correlated variable. This strategy usually works well, but it is not useful in our case because it defeats the point of the design. We are interested in the support for a fiscal package as a function of all of its individual attributes. Moreover, dropping one attribute from the analysis may lead to specification bias.

We, therefore, propose to use regularization to analyse the results from conjoint experiments, which are perfectly multicollinear. Specifically, we use ridge regression, which is a common regularization method that adds a penalty term to the common OLS regression. Hoerl (1962) and Hoerl and Kennard (1970) suggested to use a ridge regression as an ad-hoc fix to address instances of high multicollineary, including instances of design-based collinearity. It allows one to estimate coefficients for all independent variables in the model even in the presence of super collinearity, and consequently, the method is also used in fields such as genetics where a set of related genetic predictors may jointly cause certain diseases.

To see how ridge regression works, recall that OLS regression attempts to minimize the sum of errors squared, as shown in the equation below:

$$\sum_{i=1}^{M}(y_i - \hat{y}_i)^2 = \sum_{i=0}^{M}(y_i - \sum_{j=0}^{p} w_j w_{ij})^2 \tag{1}$$

Ridge regression adds the following term to this model:

$$+\lambda \sum_{j=0}^{p} w_j^0 \tag{2}$$

This term is referred to as the ridge penalty and λ is the penalty parameter. If λ is zero, then ridge regression is essentially an OLS regression. However, if λ is above zero, then it adds a constraint to the coefficient. This constraint minimizes the coefficients (which is called shrinkage), which results in a lower variance and a lower error value. Consequently, ridge regression is a way to decrease the complexity of a model without reducing the number of variables. It is a solution to a constrained estimation problem.

Importantly, the ridge penalty shrinks large regression coefficients of correlated predictors and reduces overfitting. Contrary to Lasso regression, an alternative regularization method, ridge regression does not shrink coefficients to zero.[8] It includes all independent variables in the data and is thus a good way to analyze results from conjoint survey experiments with a large number of restrictions. The "shrinkage estimators" performs better than OLS when the data matrix is relatively sparse, but it introduces a bias in the estimates due to the ridge penalty. Given that the bias introduced by the ridge penalty is systematic, it still allows us to make inferences about respondents' priorities in our case. Our empirical strategy thus provides a modelling strategy for the underlying utility function behind respondents' choice of fiscal packages, while maintaining predictive performance and interpretability.

We use the R package `glmnet` to find the best value of λ through cross-validation. We then proceed by using this optimal λ to estimate the regression coefficients (also referred to as AMCEs) and marginal means. Ridge regression does not provide standard errors for coefficients, but we rely on non-parametric boot-strapping to calculate standard errors and confidence intervals. To this end, we wrote a R function which calculates the same ridge regression 1000 times with a random sample of our observations and calculates standard errors based on the uncertainty of the results. This is a popular method for parametric inference, and it allows us to assign a measure of accuracy to the coefficients obtained from the ridge regression.

[8] Lasso regression is thus especially useful for model selection.

E Analysis by country

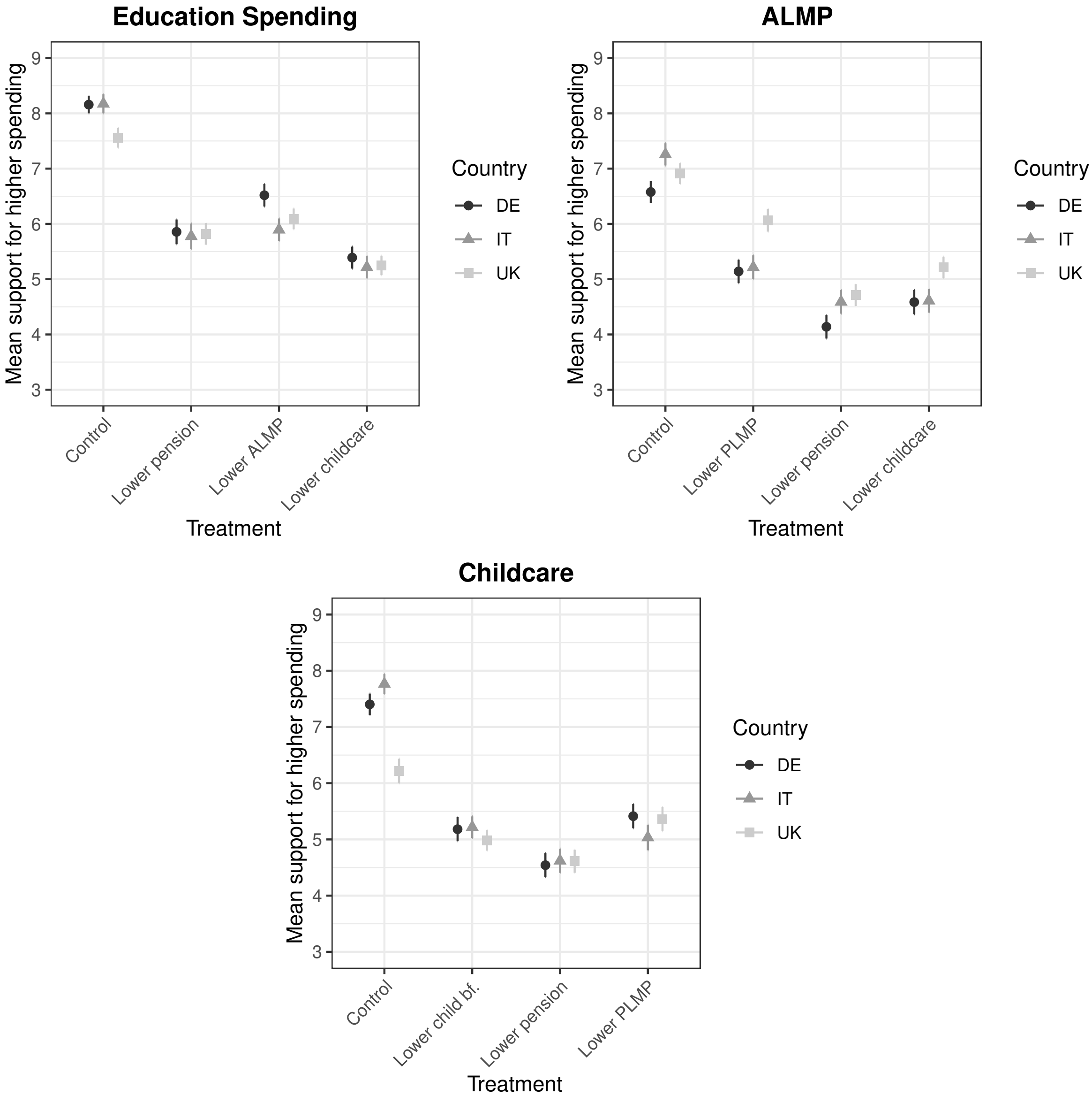

Figure A.3: Mean support for spending increases with and without trade-offs, by country

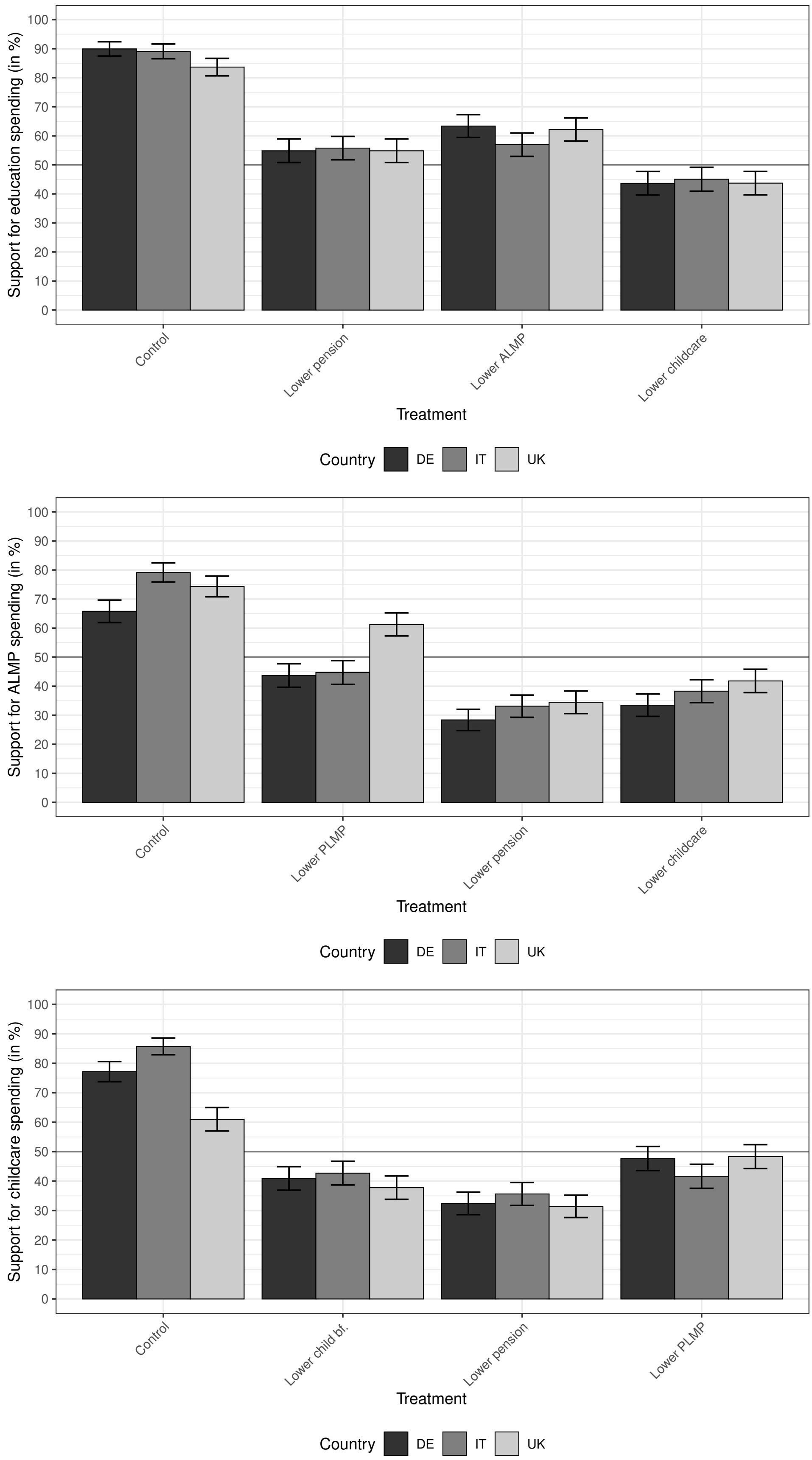

Figure A.4: Share of supporters for spending increases by treatment, by country

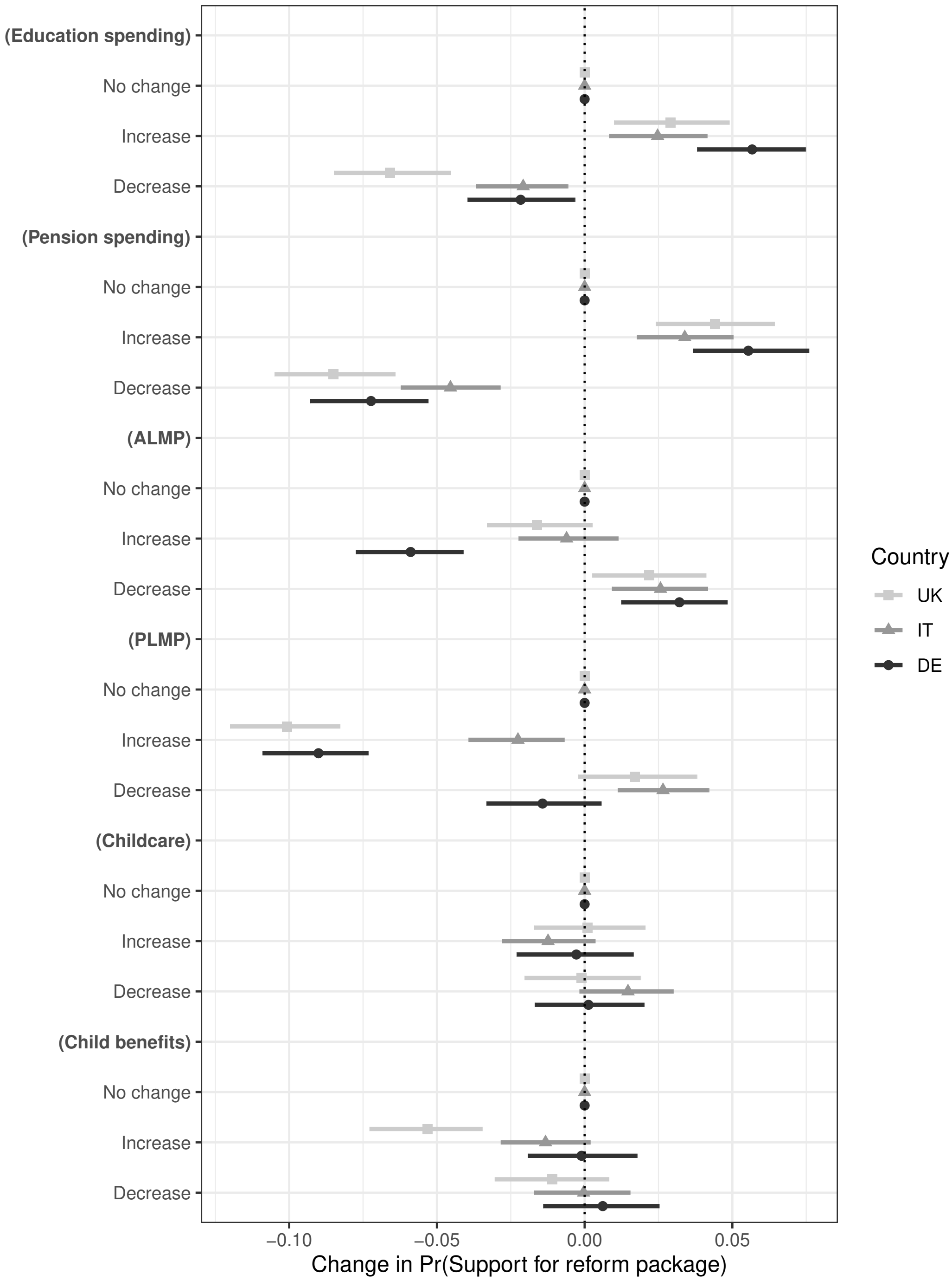

Figure A.5: AMCEs from conjoint survey experiment by country

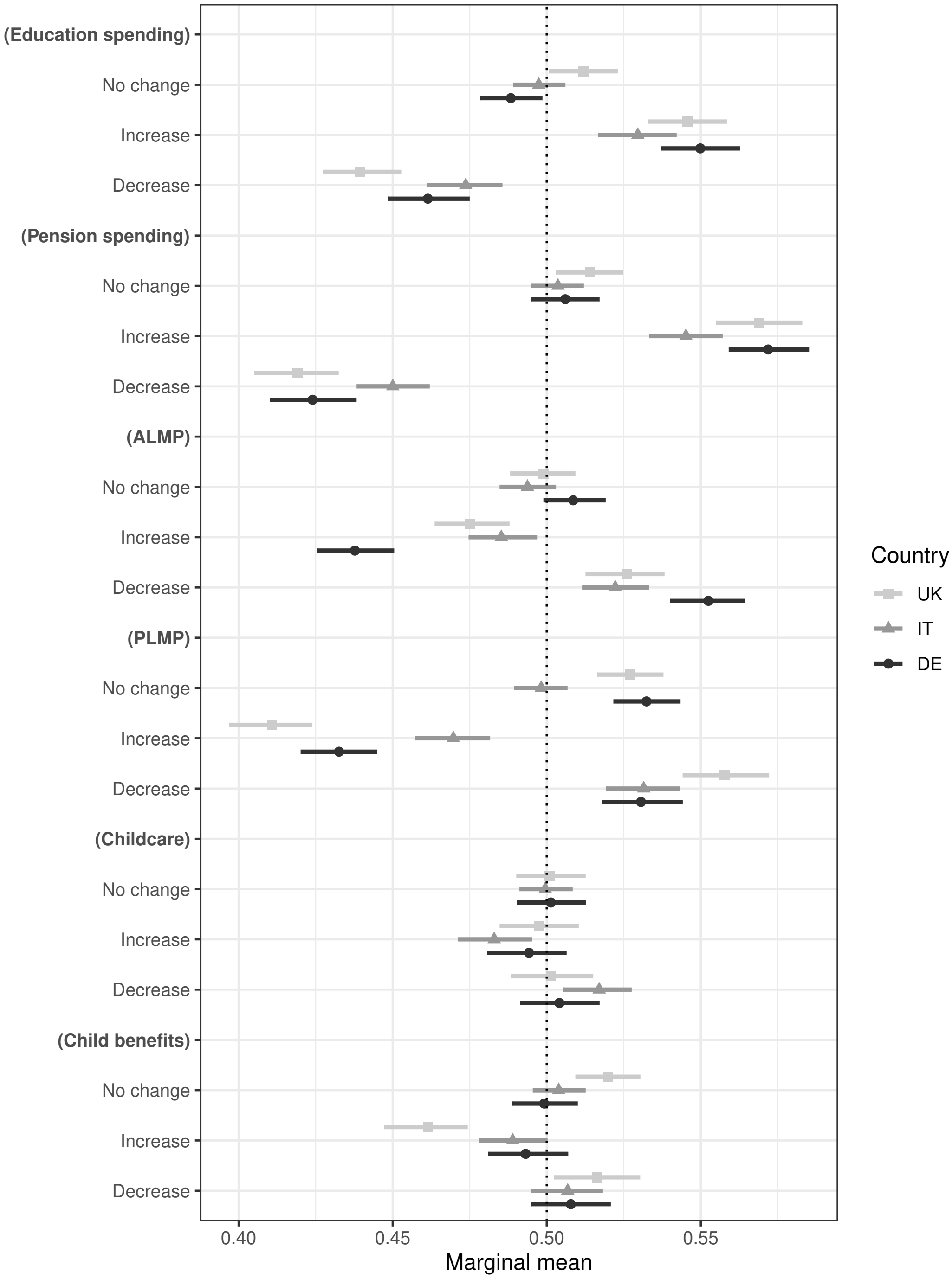

Figure A.6: Estimated marginal means from conjoint survey experiment by country

F Analysis with weights created by entropy balancing

We worked with the survey company Qualtrics to obtain an online sample that was representative of the population based on key indicators such as age and gender in each country. To test that there were no other biases in our sample that influence our results, however, we used entropy balancing to create survey weights using the R package `ebal`. Based on margins from the population in each country, we created these weights based on age, gender, education, and economic activity and replicated our analyses. The results are not substantively different from the results shown in the main text.

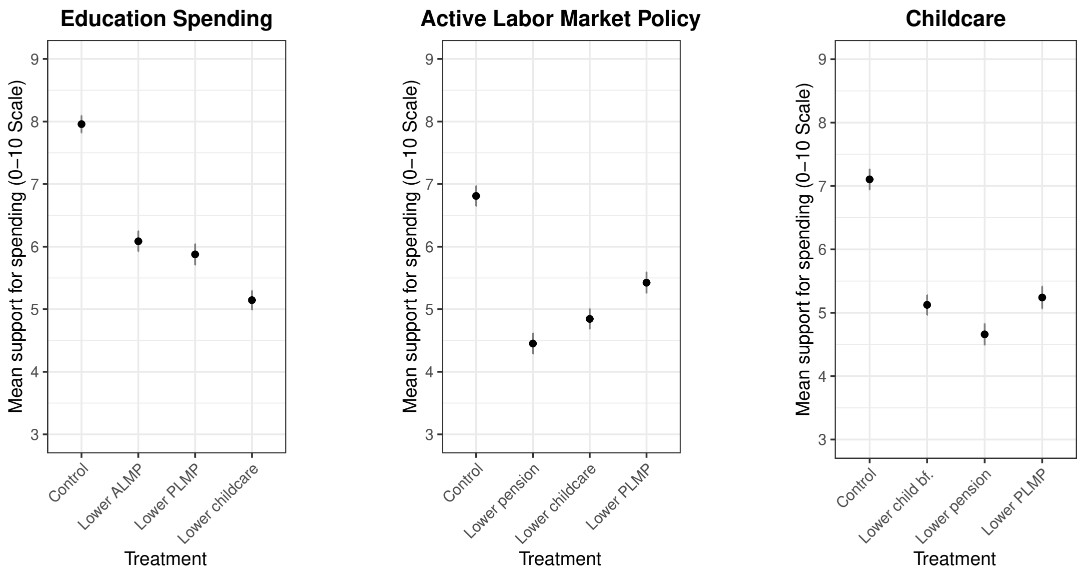

Figure A.7: Average support for spending increases by treatment with survey weights, pooled
Note: The figure shows the mean support and 95 percent confidence intervals for different policies without trade-offs (control group) as well as with the three respective trade-offs (treatment groups). Mean support is measured on a scale from 0 to 10.

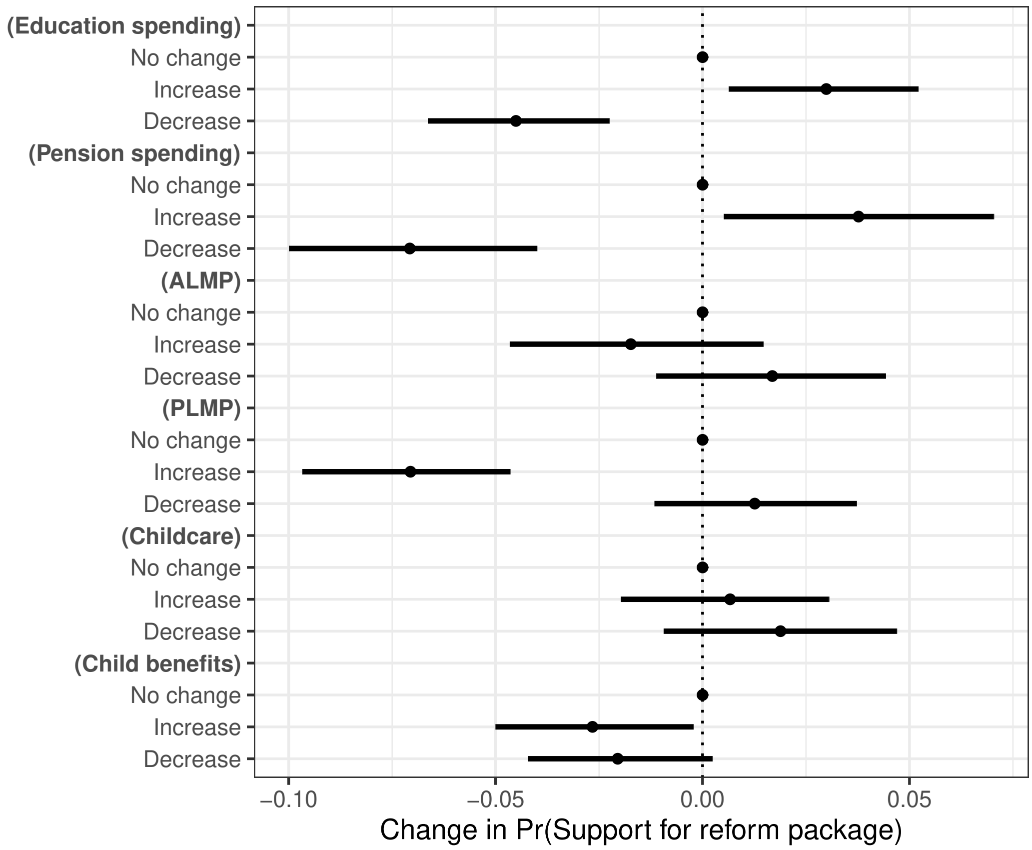

Figure A.8: AMCEs from conjoint survey experiment with survey weights

G Additional results from the conjoint survey experiment

We replicate our analysis with a second dependent variable, namely the rating variable. For this variable, respondents evaluated each reform proposal that they saw on a scale from 0 to 10. The results are very similar to the findings presented in the main text. We then show the full conjoint graphs for the difference between constituencies and non-constituencies shown in Figure 5. Moreover, to test the robustness of our findings, we further run several tests for heterogeneous effects in our conjoint survey experiment. Specifically, we tested for heterogeneous effects along for subgroups by ideology (left-right self placement), education level, and income. The results are shown below and they indicate that the heterogeneous effects are relatively small, except for specific groups on specific attributes (see our discussion on policy constituencies in the main text). This is also true for other possible subgroups (e.g., employment activity, gender, age) which are not shown below.

G.1 Results from the analysis with the rating variable

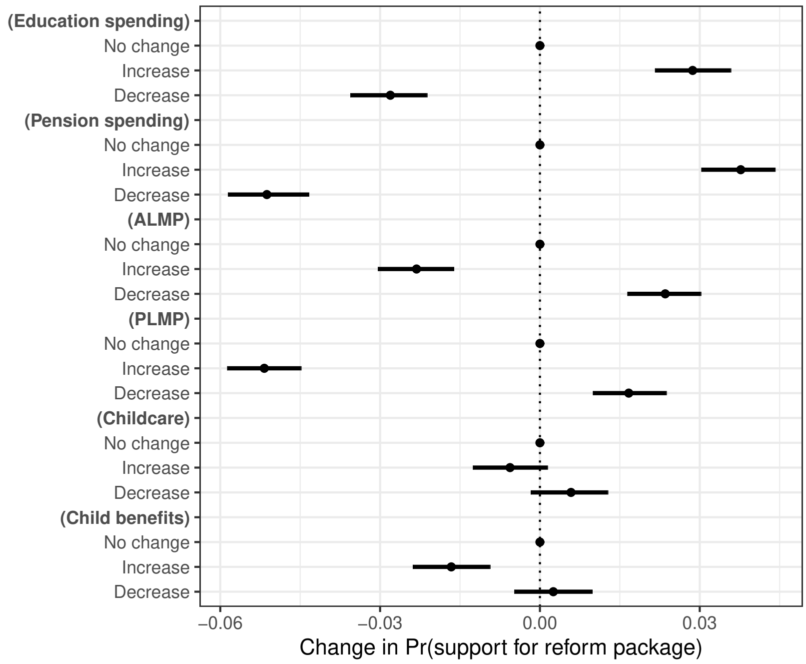

Figure A.9: AMCEs from conjoint survey experiment with the rating variable, pooled
Note: The figure shows the average component-specific marginal effect (ACME) of a change in the value of one of our six dimensions on the support for the reform package.

G.2 Full constituency plots showing heterogeneous effects

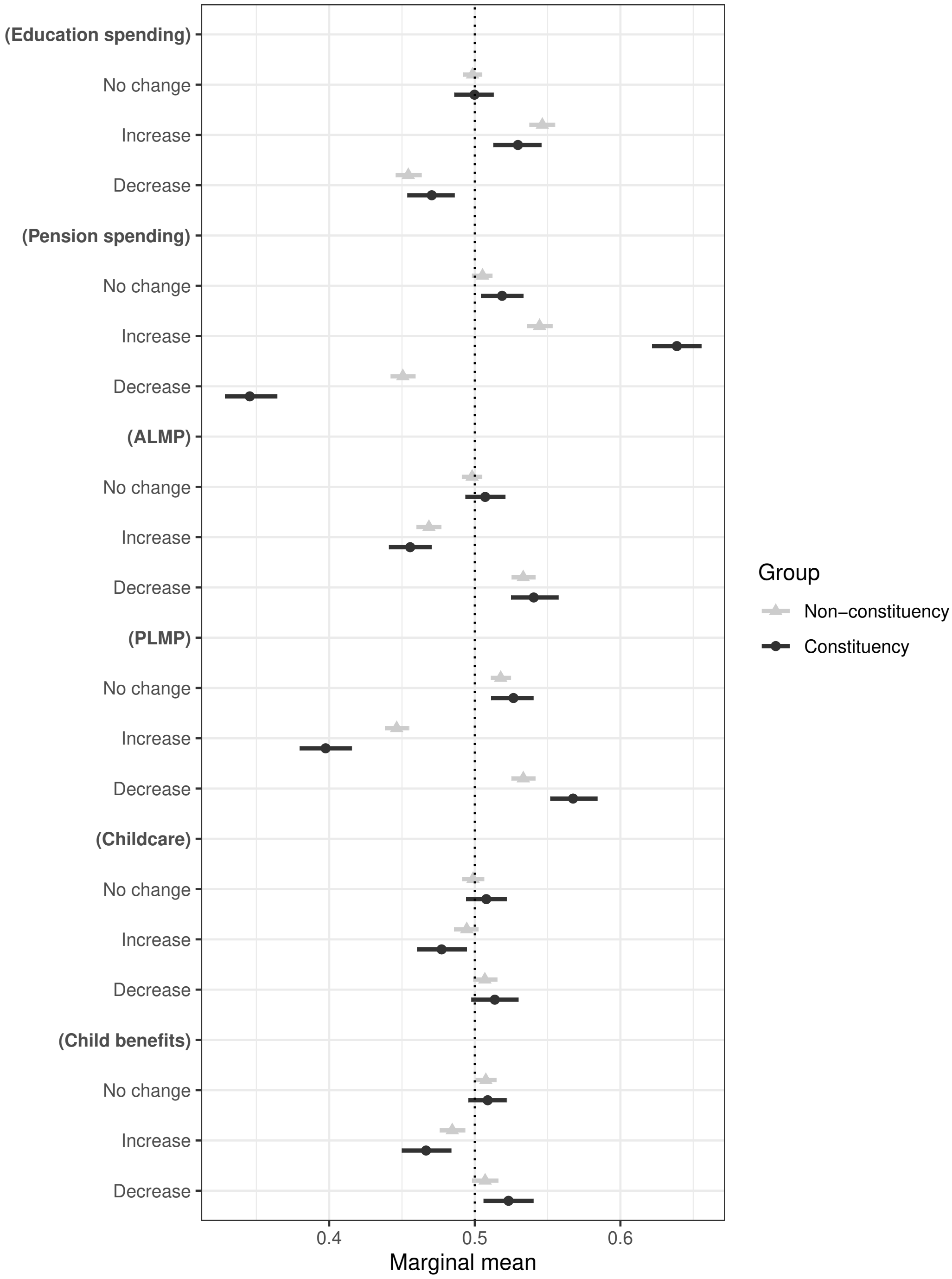

Figure A.10: Estimated marginal means from conjoint survey experiment for retired respondents vs. and all other respondents

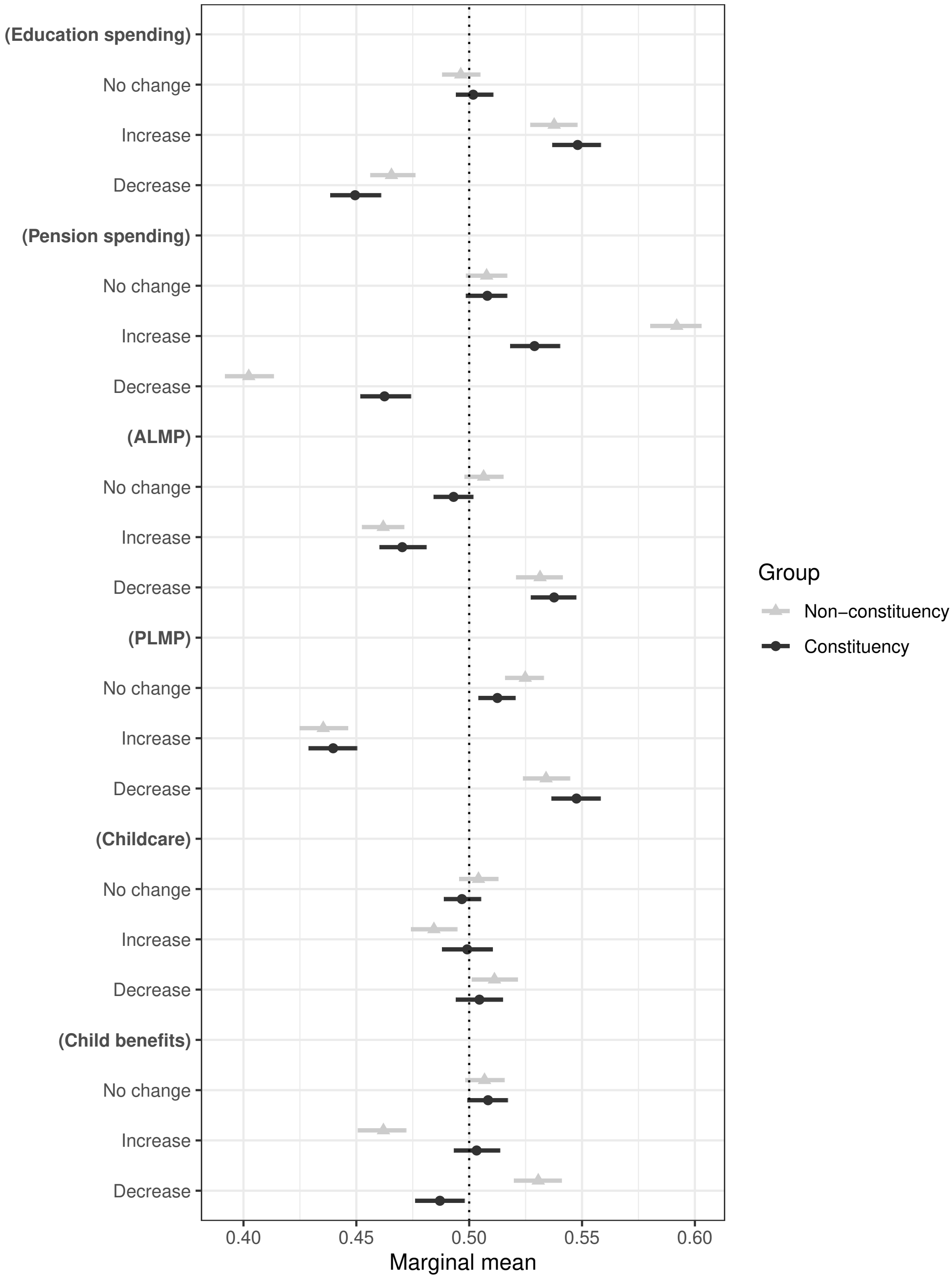

Figure A.11: Estimated marginal means from conjoint survey experiment for education constituency vs. all other respondents

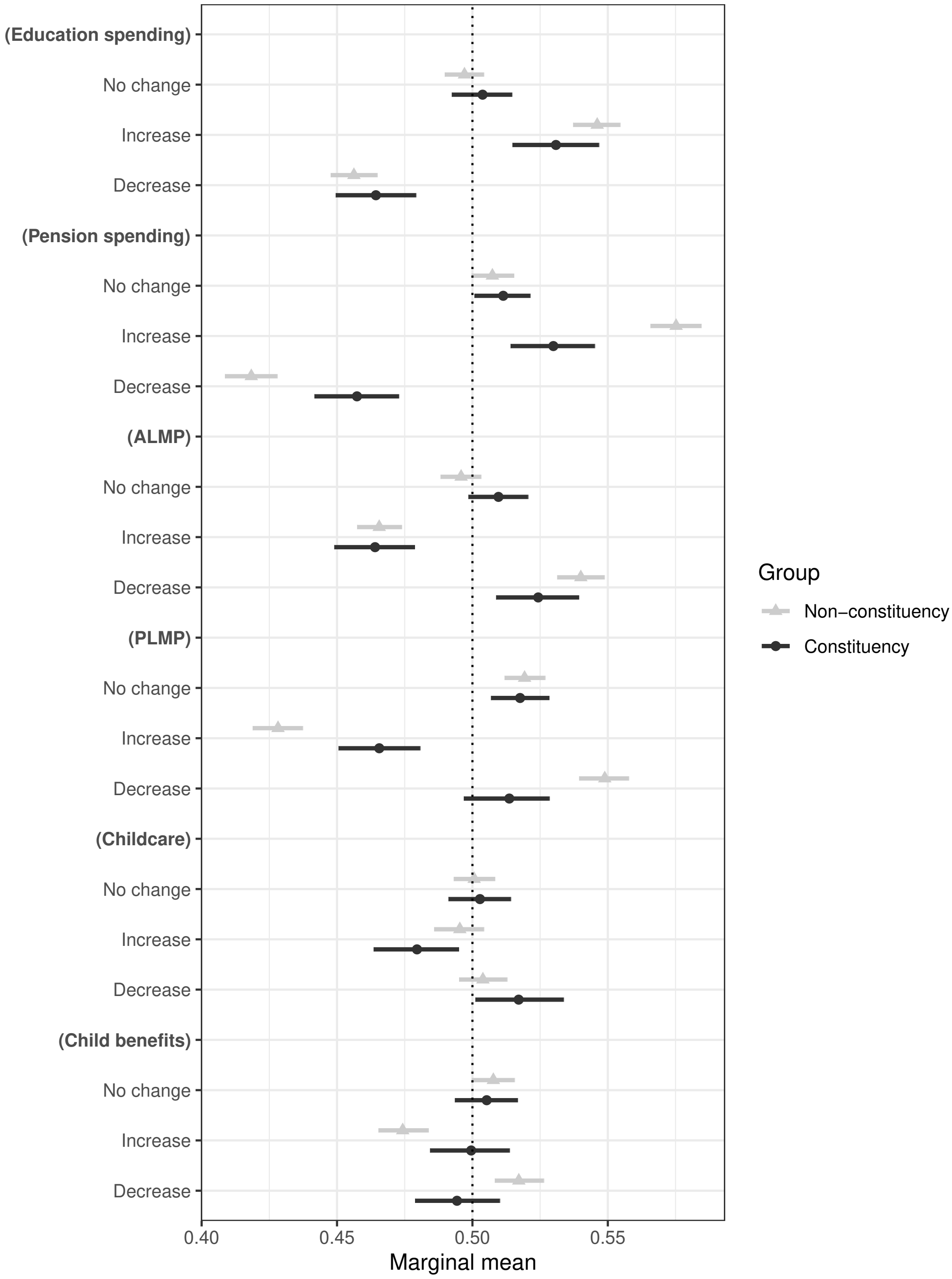

Figure A.12: Estimated marginal means from conjoint survey experiment for outsiders vs. all other respondents

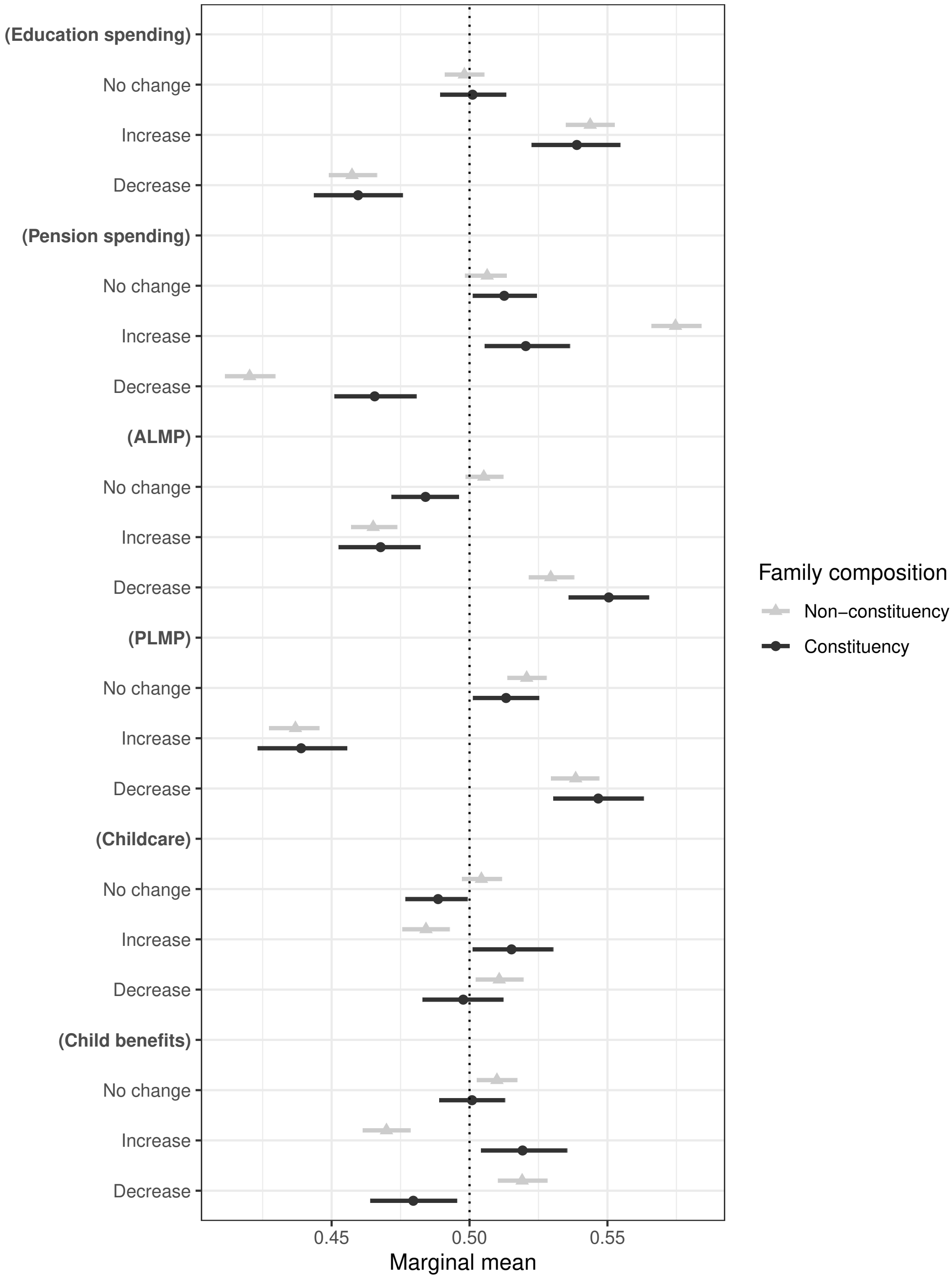

Figure A.13: Estimated marginal means from conjoint survey experiment for parents of children (under the age of 10) vs. all other respondents

G.3 Additional tests for heterogeneous treatment effects

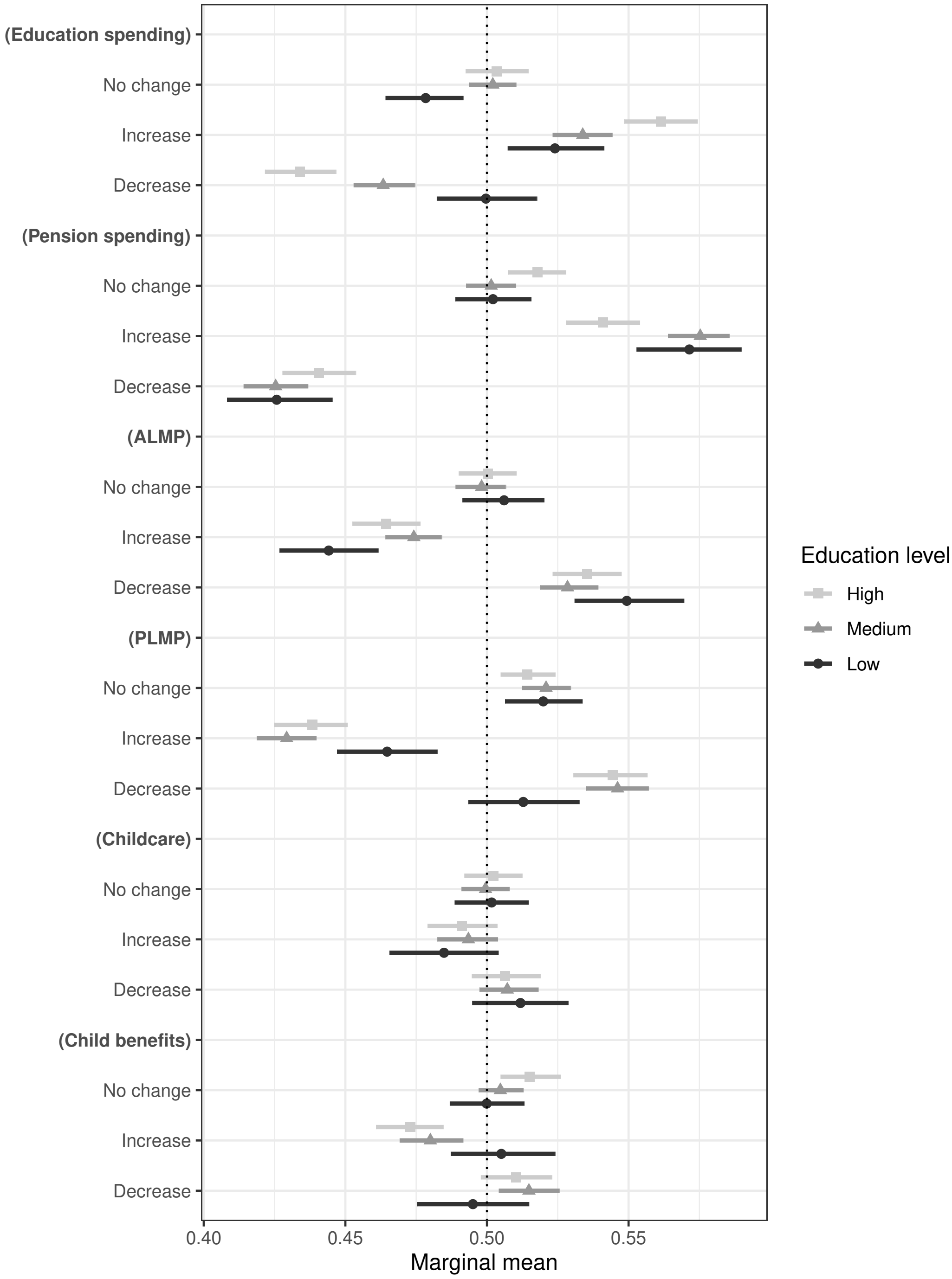

Figure A.14: Estimated marginal means from conjoint survey experiment by education

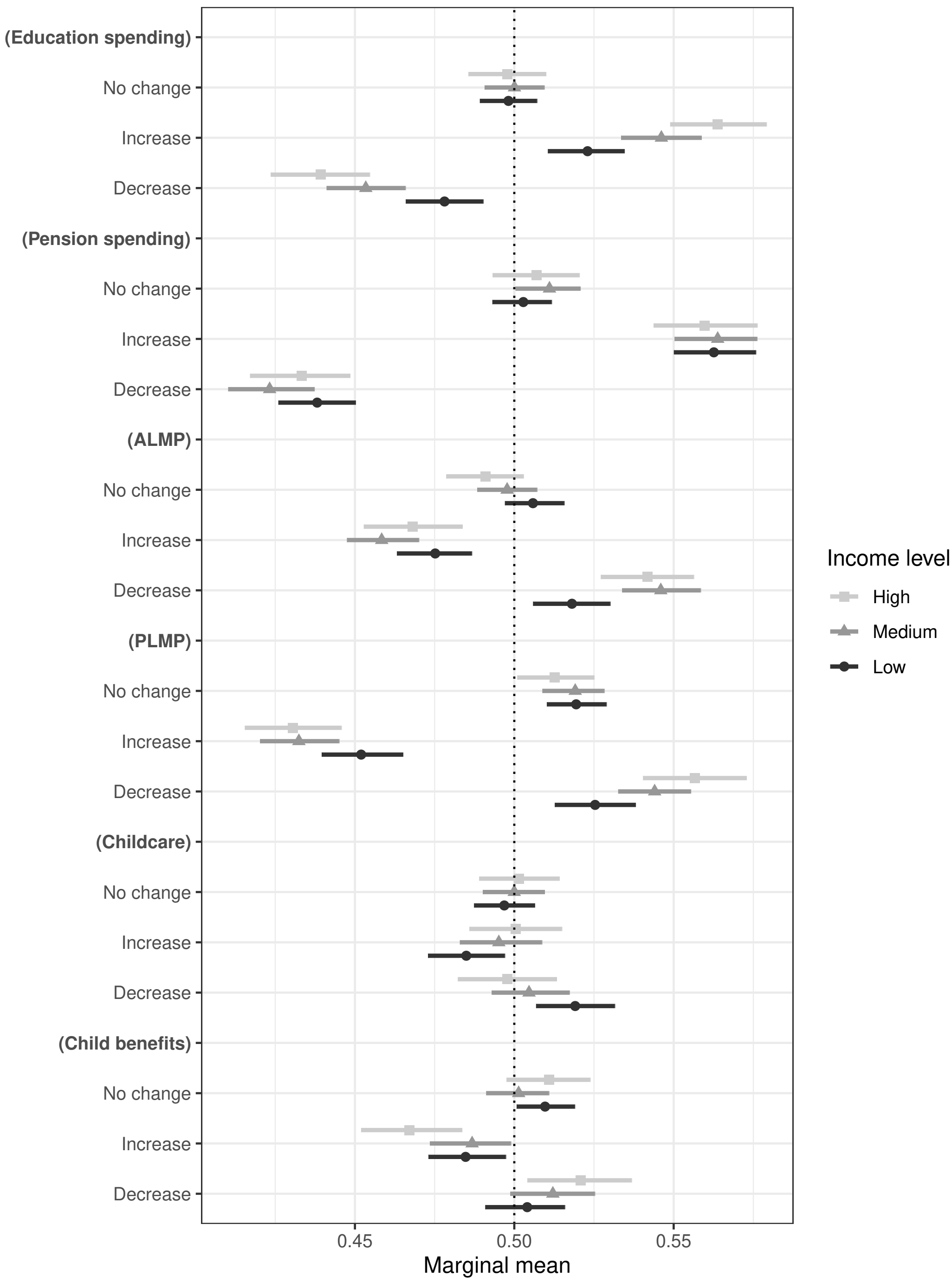

Figure A.15: Estimated marginal means from conjoint survey experiment by income

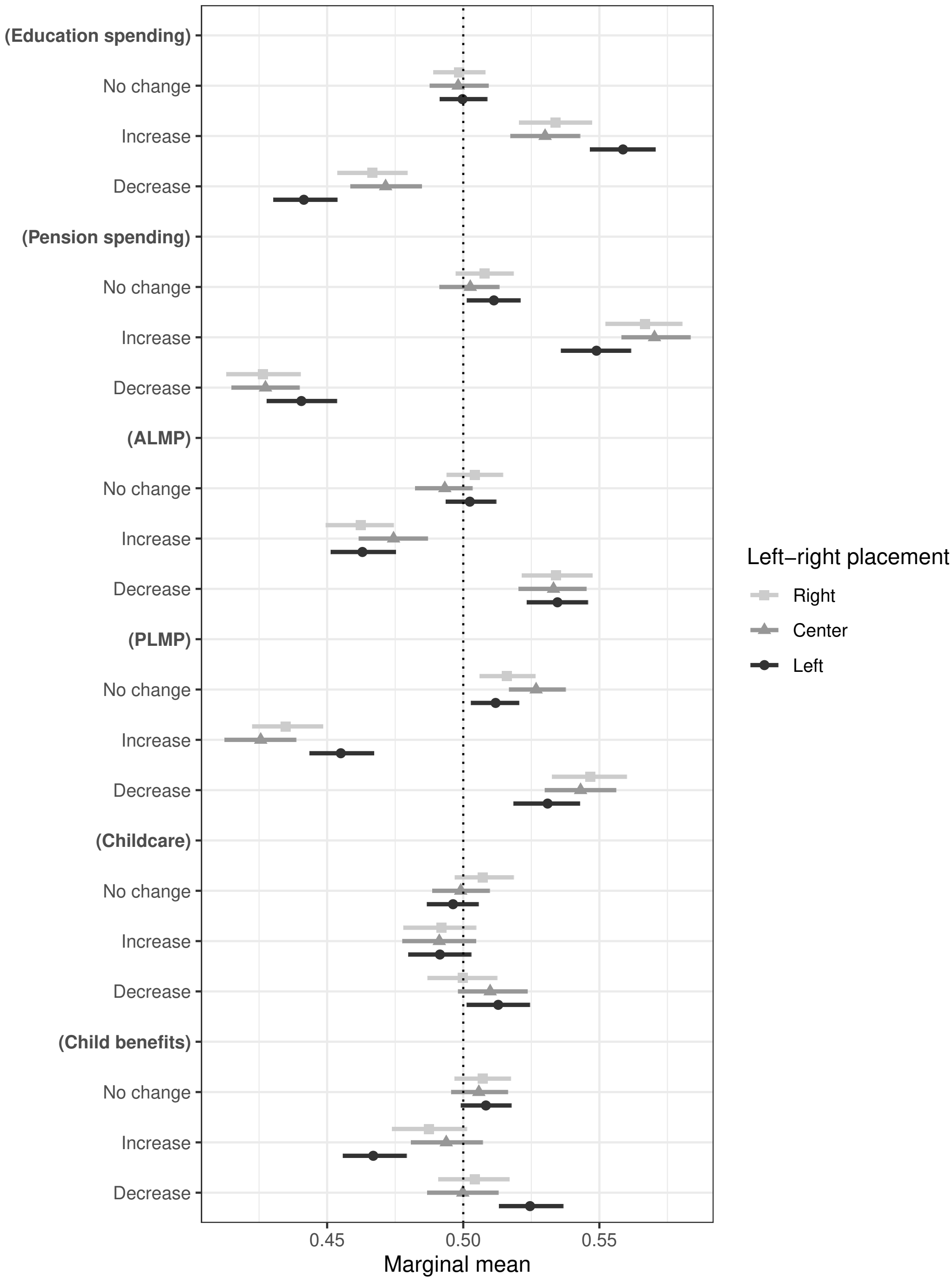

Figure A.16: Estimated marginal means from conjoint survey experiment by left-right placement

H Robustness tests for the conjoint survey experiment

We have used a series of tests to check whether the results are robust. They were designed to check that the common assumptions involved in conjoint analysis are satisfied and to probe potential concerns about the validity of the results.

On the one hand, we conducted the diagnostic tests suggested by Hainmueller et al. (2014). First, conjoint analyses relies on the assumption that there are no carryover effects between the different rounds of conjoint tasks. To test whether this assumption holds, we estimated the AMCEs separately for each of the five rounds of conjoint tasks. The results are shown in figures A.17.[9] Second, we checked whether there are profile order effects, i.e., whether the AMCEs and marginal means depend on whether the attribute occurs in the first or second profile in a given task. To this end, we estimated AMCEs separately for all the observations where attribute levels occurred in the first and the second profile respectively (Figure A.18). Finally, note that we already addressed the concern about atypical profiles raised by Hainmueller et al. (2014) in the research design. Specifically, we included a large number of restrictions to prevent profiles that are unrealistic and would not occur in the real world.

On the other hand, we also used further robustness tests, which are important due to the design of the survey. First, we checked whether respondents lost concentration throughout the survey by estimating all results based on the first two (out of five) conjoint comparisons only. Moreover, we included round or task fixed effects to take account of the fact that respondents might make different choices in later stages of the conjoint experiment, for example due to fatigue or lack of concentration. The results for the first two (out of five) conjoint comparisons are shown in Figure A.19.

Second, we assessed the relative time that respondents took to complete the conjoint tasks and we excluded those respondents that speed through the conjoint tasks, comparing the results with the overall sample. We also distinguished respondents by the time that they took overall for the survey and used subgroup analysis to test whether our results are robust across groups. The results are shown in Figure A.20.

Third, the conjoint survey experiment described above was embedded in a survey, which included two different set of conjoint tasks. The order in which these conjoint experiments occurs in the survey was randomized. Still, we checked whether respondents are influenced in their evaluations of the conjoint profiles if they have already completed a different set of conjoint tasks beforehand. For this purpose, we split the sample and analyzed the results separately depending on whether the conjoint experiment occurred before or after the other conjoint experiment in the survey (Figures A.21).

[9] Following, the suggestion by Leeper et al. (2020), we also estimated the conditional marginal means for each of the five rounds of conjoint tasks but the results are not shown.

Fourth, there is also a possibility that the screen size might affect the way respondents evaluate the conjoint tasks. We therefore also separately analyzed responses from mobile versus non-mobile respondents and checked to what extent they differ.

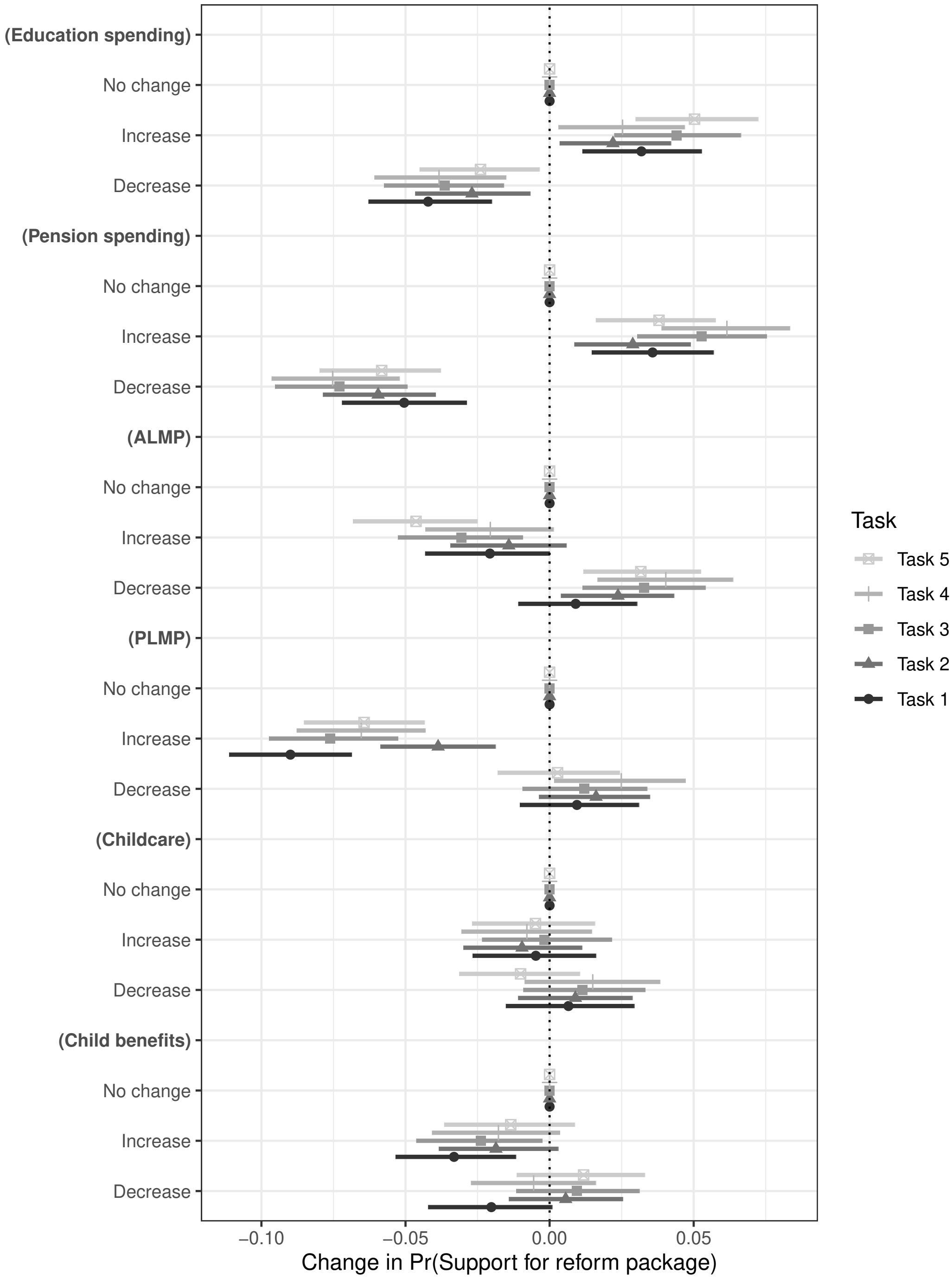

Figure A.17: AMCEs from conjoint survey experiment by task

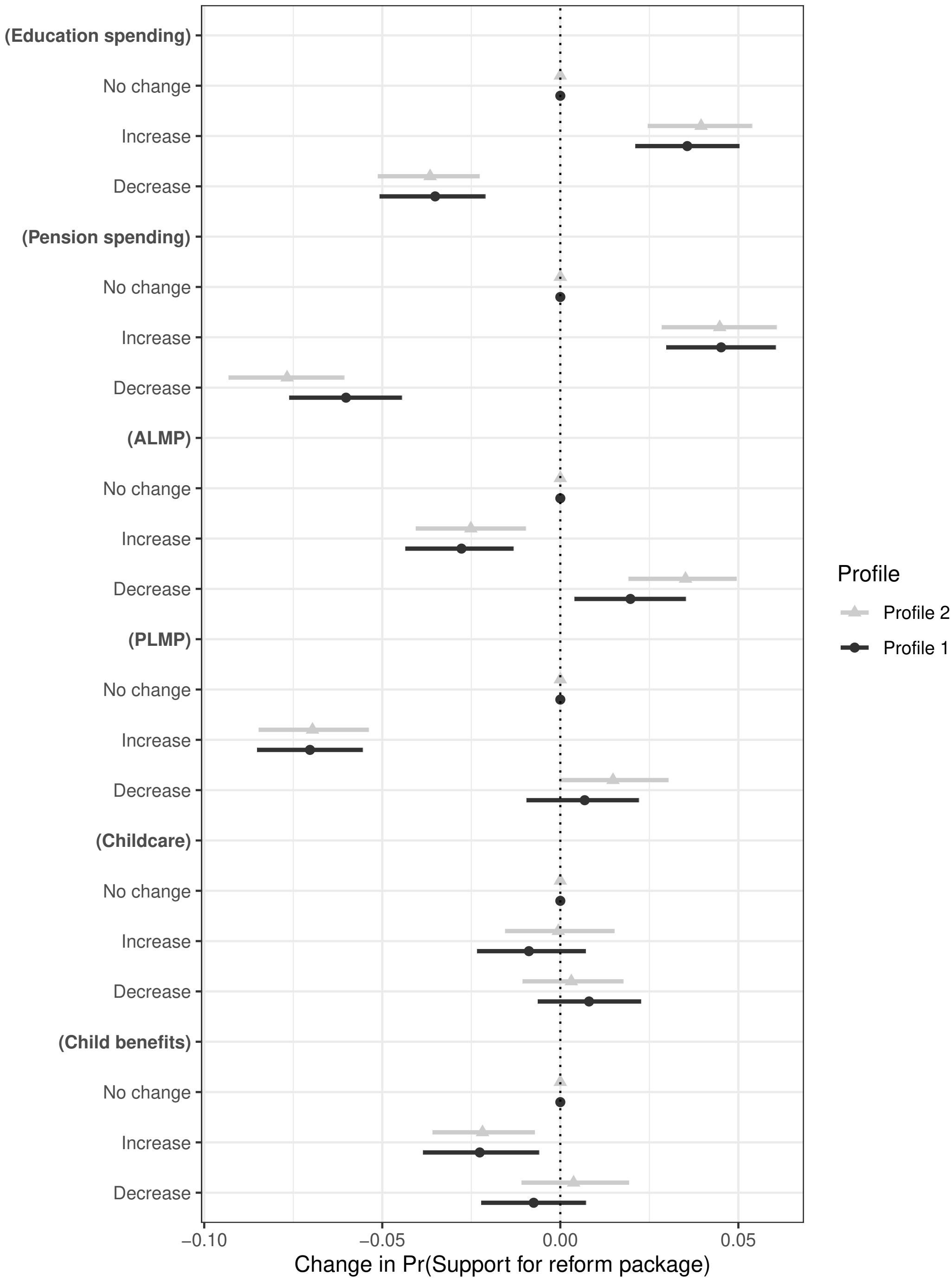

Figure A.18: AMCEs from conjoint survey experiment by profile order

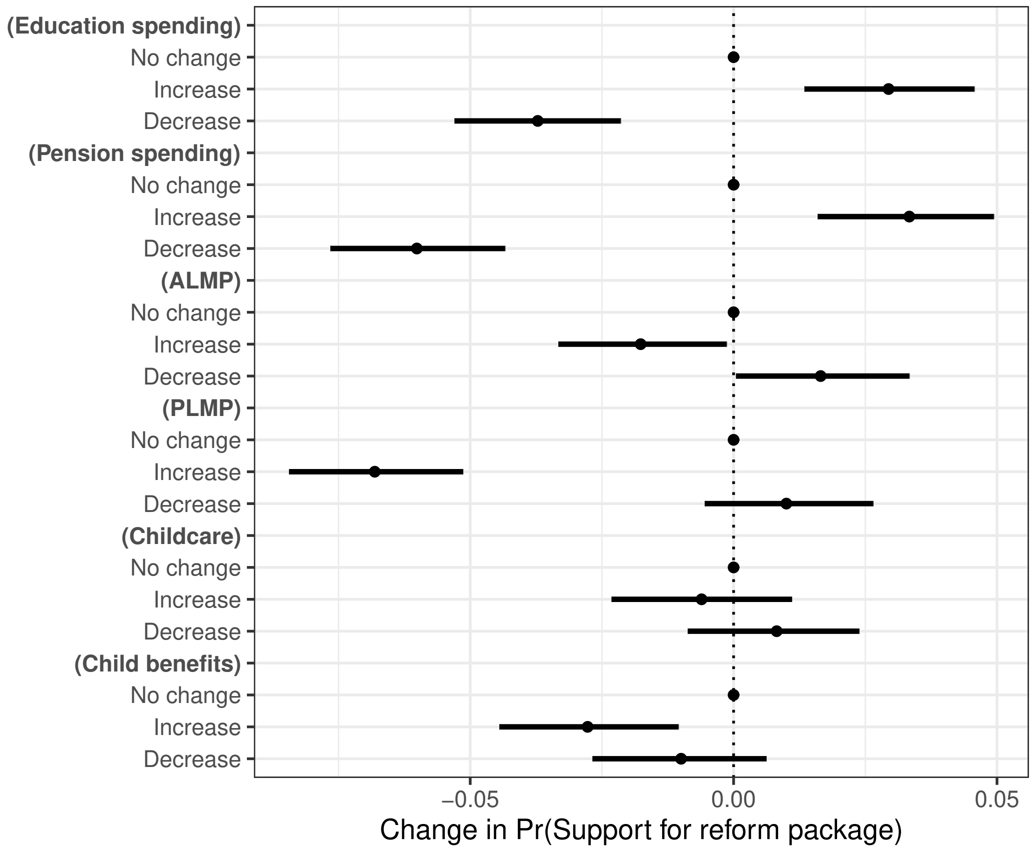

Figure A.19: AMCEs from conjoint survey experiment for the first two tasks only

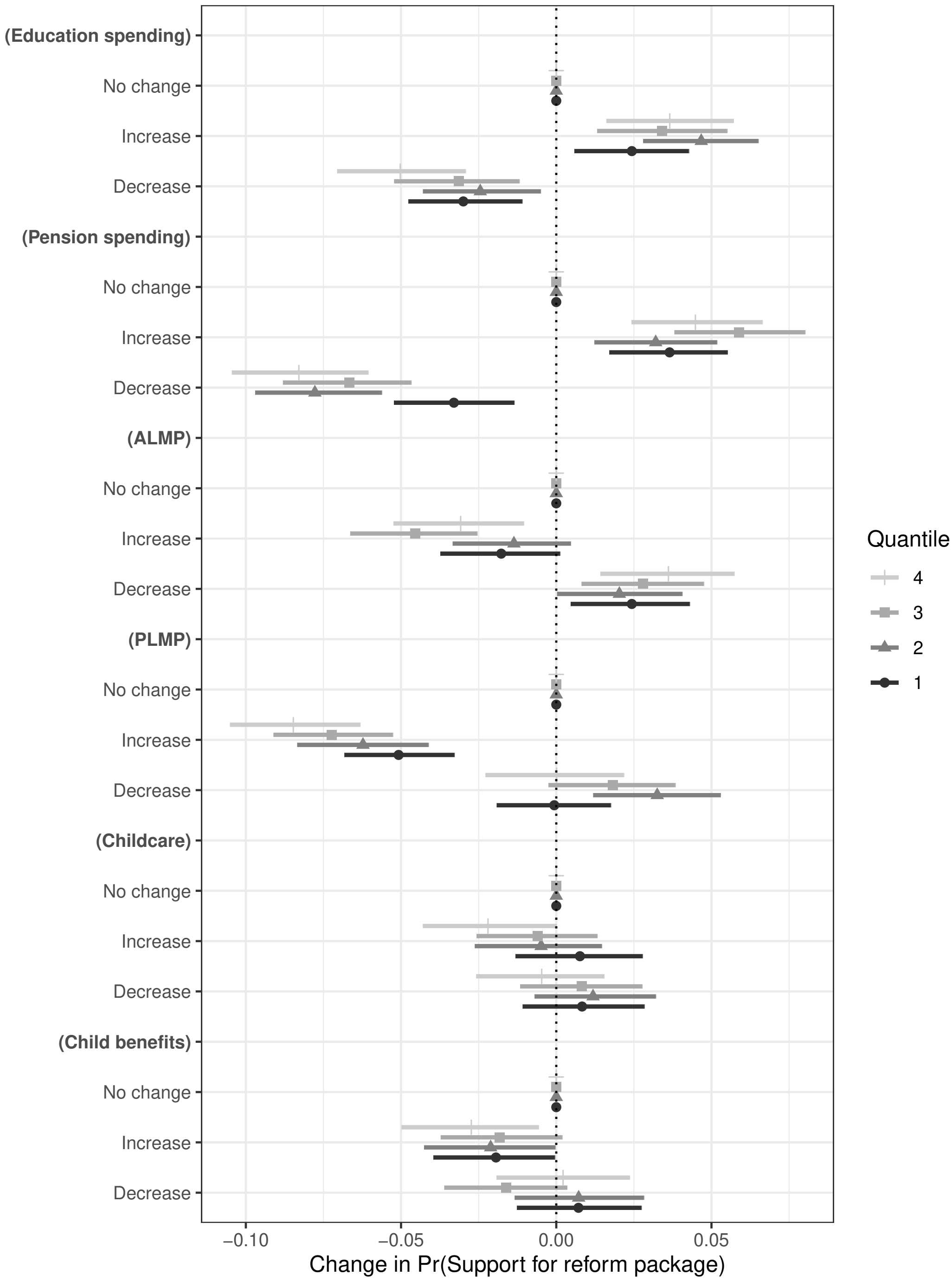

Figure A.20: AMCEs from conjoint survey experiment by speed of respondents

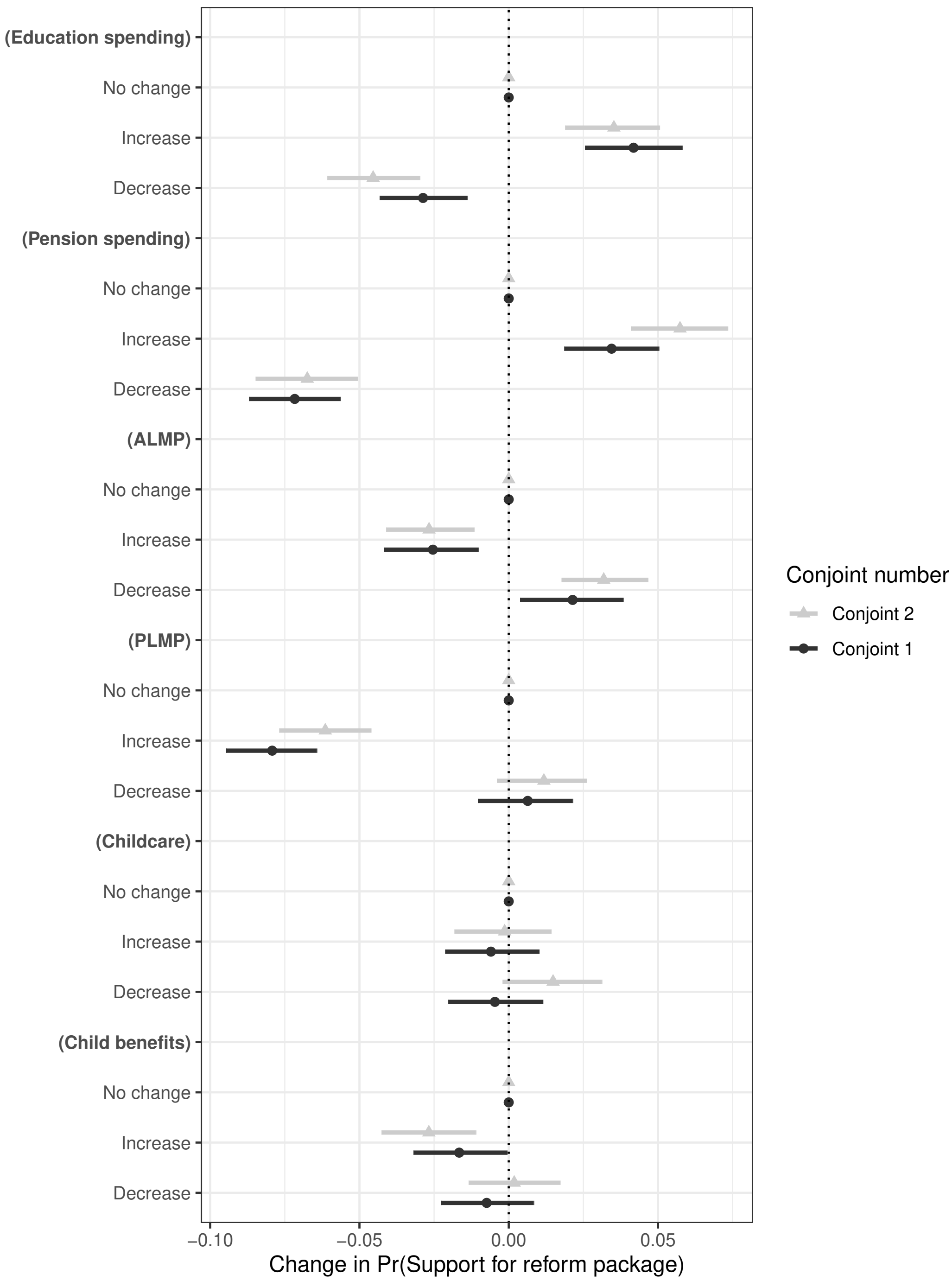

Figure A.21: AMCEs from conjoint survey experiment by order of the conjoint task in the survey

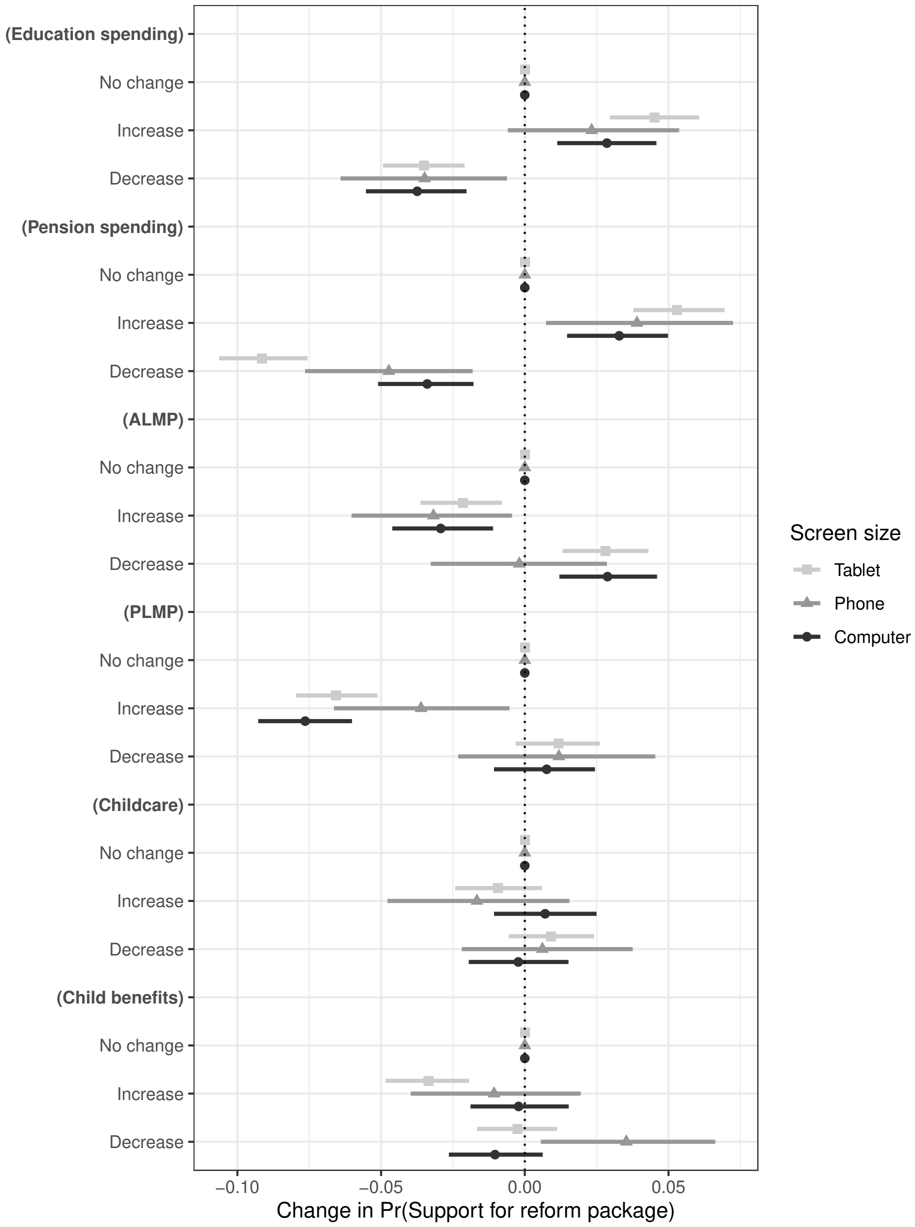

Figure A.22: Estimated marginal means from conjoint survey experiment by screen size of respondent's device