How and why to Involve Policy Makers More Structurally Into Impact Evaluations

David McKenzie, *World Bank*

(joint work with Leonardo Iacovone and Rachael Meager)

The big picture question: how can we learn better from pilots and government programs?

- Policymakers/researchers often run pilots and programs that cost considerable time and money, but with limited sample sizes
    - Especially true with programs for SMEs
        - Iacovone et al. (2021) : 159 autoparts firms in Colombia, US$2.4 million, 4 years
        - Bruhn et al. (2018): 150 firms treated in Puebla, 4-5 years
        - Higuchi et al. (2017) – 312 firms Vietnam, divided into 4 groups, tracked for 5 years
        - Custodio et al. (2021) – 93 firms in Mozambique, tracked for a year
- Budget plus limits on possible sample present challenges for statistical power
- If find $p>0.10$, can’t reject program had no effect, but then what should government do, or what should researchers conclude?

Other common results/policy interpretations of standard impact evaluation

- Researchers find no impact on outcome Y1 (e.g. exports)
  - Policymakers respond – but of course, we never expected it to affect Y1 anyway, the program was really designed to affect Y2 (e.g. productivity)
- Researchers find significant impact on outcome Y1
  - Of course, this is what we knew all along, this impact evaluation didn't really teach us anything
- Researchers find a negative impact on outcome Y1
  - Well, I still believe this program works, we must have just got unlucky in this small sample.

Our proposal: formally incorporate policymaker information in Bayesian analysis

- Interventions don’t occur in a void
  - Policymakers have experience, beliefs about what program will do
  - Researchers have theory, existing body of evidence
  - Participants themselves apply for such programs based on belief about how much it will help them
- Given the results of the experiment, how much should they update their beliefs about the effectiveness of this type of intervention?
  - Bayesian analysis provides a way

Illustrate how to do this through a real policy application in Colombia

- *Substantive policy issue:* diversifying and expanding the export base a key policy objective – Colombia highly dependent on a few commodities (petroleum, coal, coffee, flowers)
- *Colombian program* aims to broaden range of firms and sectors engaging in exporting
  - aims to do so through improvements in management practices coming from individualized technical assistance
- Government selected 200 firms for the pilot
  - 100 treated get diagnostic + 190 hours of technical assistance, at cost of approximately $14,000 per firm.
  - 100 control just get diagnostic
- Ex-ante power calculations suggest reasonable power to detect improvements in business practices and in binary outcomes like whether firms export, but low power/high MDEs for skewed continuous outcomes like export value.

Step 1: get agreement on which outcomes program is meant to affect, how you will measure them, and over what time frame.

- Makes it really precise what success means, and also helps make sure you are measuring what really matters, and at the right time for the program.
- E.g. Program impact on:
  - Whether firms export or not in the year after the intervention compared to the control group
  - Export diversity: the number of country-product combinations they export in the year after the intervention compared to the control group
  - Growth in export value: the percentage higher export value will be in the treatment group than control group on average in the year after the intervention

Step 2: Elicit priors from policymakers (and maybe researchers and firms)

- Helpful to do this after baseline/application data are available, so you can be clear what the firms look like, explain the intervention clearly, and give a sense of baseline values.
- Don’t just want an estimate of the predicted effect, but also the uncertainty around this effect.
  - Use a bins and beans or other type of approach
- Collect priors from multiple policymakers involved in program decisions and average these out to use wisdom of crowds.
- Fit a distribution to these priors.
- Also ask what threshold would be used to determine whether the program should be continued/expanded etc.

This outcome combines the number of distinct products with the number of countries. For example, a firm that exports cotton t-shirts to Brazil and Argentina will have 2 product-country combinations, as would a firm that exports both gulupa (purple passionfruit) and pitaya (dragonfruit) to the United States. In 2017, on average the 100 firms offered the full intervention exported 9.8 different country-product combinations. This is the average over both exporters and non-exporters, so includes zeros for the half of the firms that do not export.

We want to know how much you think this will change for the group getting offered the full intervention compared to getting offered just the diagnostic and trade fair, over the first 12 months since firms start their implementation. For example, if you think there is a 30 percent chance the intervention will increase the average number of product-country combinations firms export to by 3.3 to 3.7 product-countries, put 6 stones in the box 3.3 to 3.7, and allocate your remaining stones according to what else you think is likely.

BELIEFS ABOUT THE IMPACT ON THE NUMBER OF PRODUCT-COUNTRIES THAT FIRMS EXPORT TO

-10.1 or less	-10 to -8.1	-8 to -7.3	-7.2 to -6.8	-6.7 to -6.3
-5.2 to -4.8	-4.7 to -4.3	-4.2 to -3.8	-3.7 to -3.3	-3.2 to -2.8
-2.7 to -2.3	-2.2 to -1.8	-1.7 to -1.3	-1.2 to -0.8	-0.7 to -0.3
-0.2 to 0.2	0.3 to 0.7	0.8 to 1.2	1.3 to 1.7	1.8 to 2.2
2.3 to 2.7	2.8 to 3.2	3.3 to 3.7	3.8 to 4.2	4.3 to 4.7
4.8 to 5.2	5.3 to 5.7	5.8 to 6.2	6.3 to 6.7	6.8 to 7.2

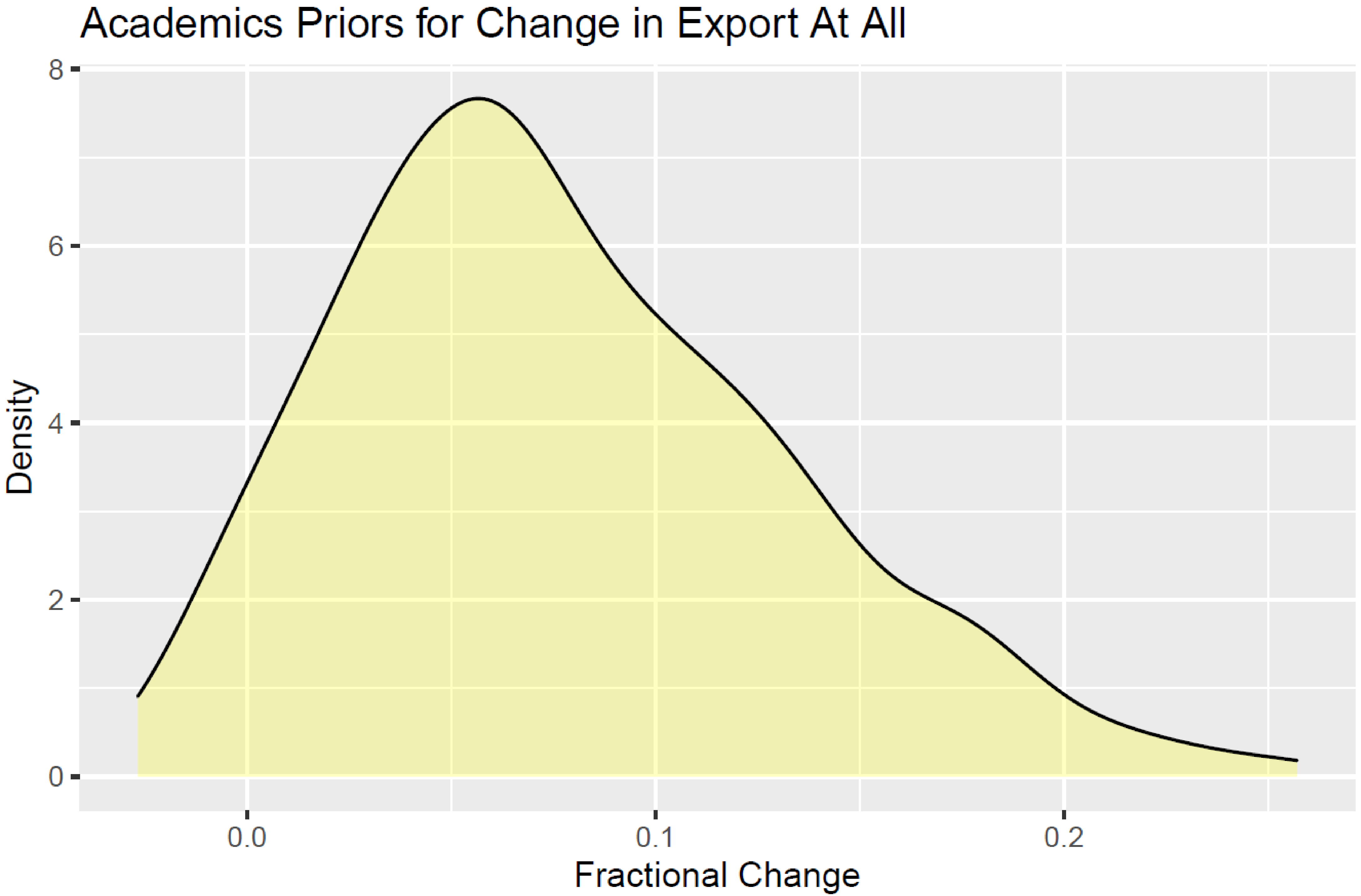
Step 3: conduct the experiment and collect data to compare treatment and control

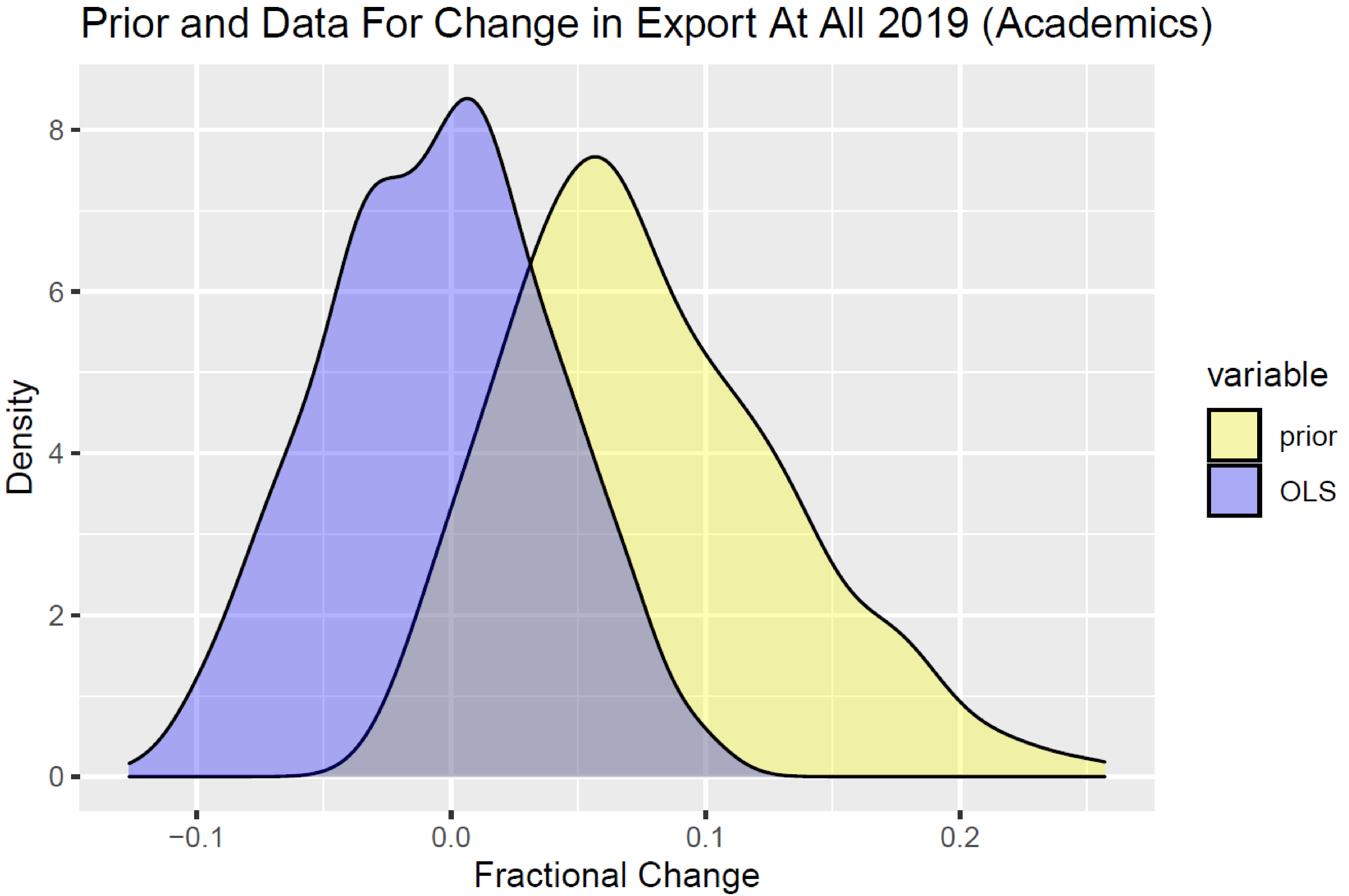
- This would be the standard (frequentist) analysis.
- E.g. collect data on firm exports in year after the intervention

Step 4: Use Bayesian analysis to update the prior information with the data, to obtain posterior probabilities

Step 5: also use the fitted posteriors for decision analysis

e.g. what is the probability the impact of the program was as much as needed to pass cost-effectiveness?


Academics Priors for Change in Export At All
Density
8
6
4
2
0
0.0
0.1
0.2
Fractional Change


Prior and Data For Change in Export At All 2019 (Academics)
Density
8
6
4
2
0
−0.1
0.0
0.1
0.2
Fractional Change
variable
prior
OLS

Bayesian Inference On Change in Export At All 2019 (Academics)

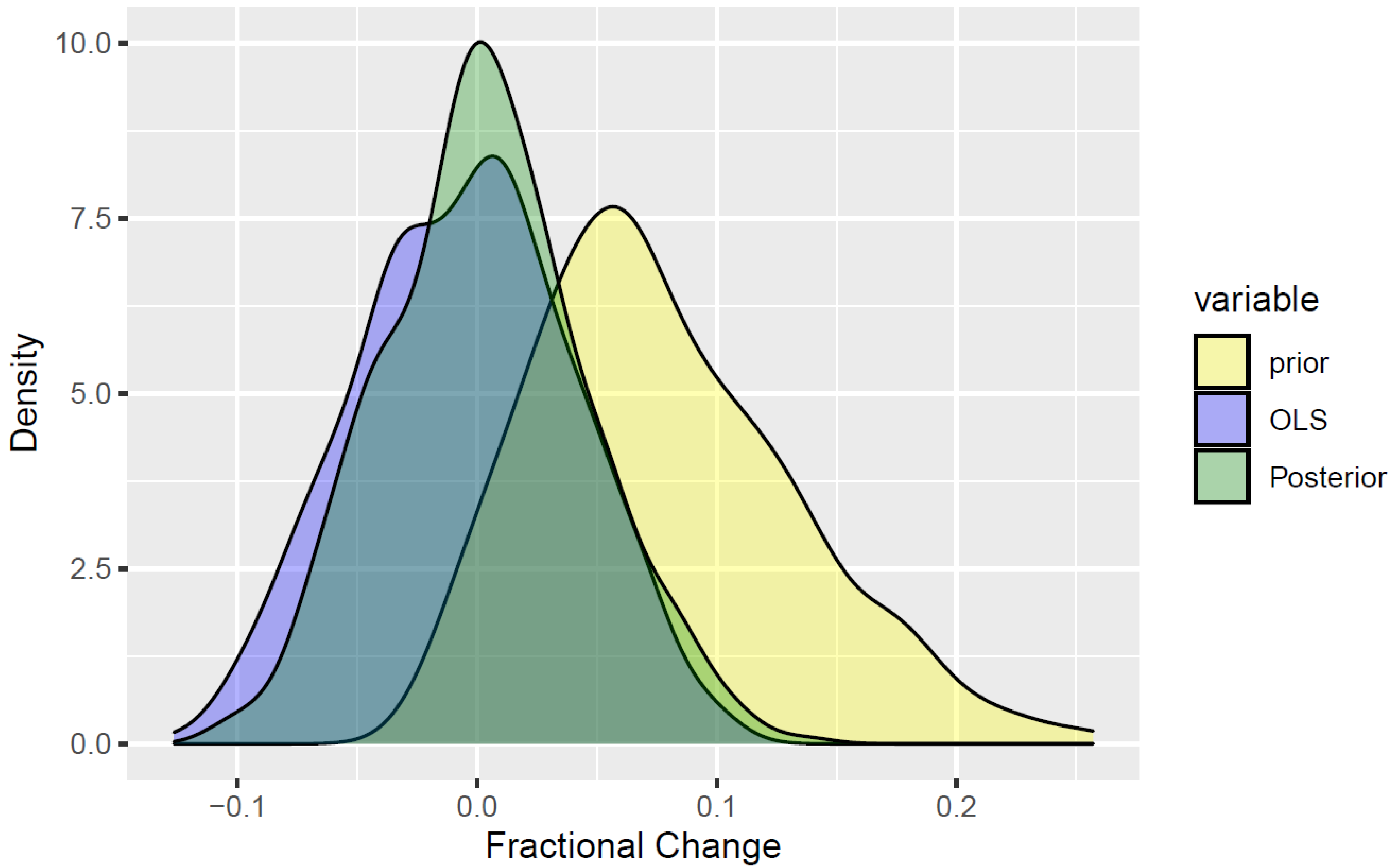
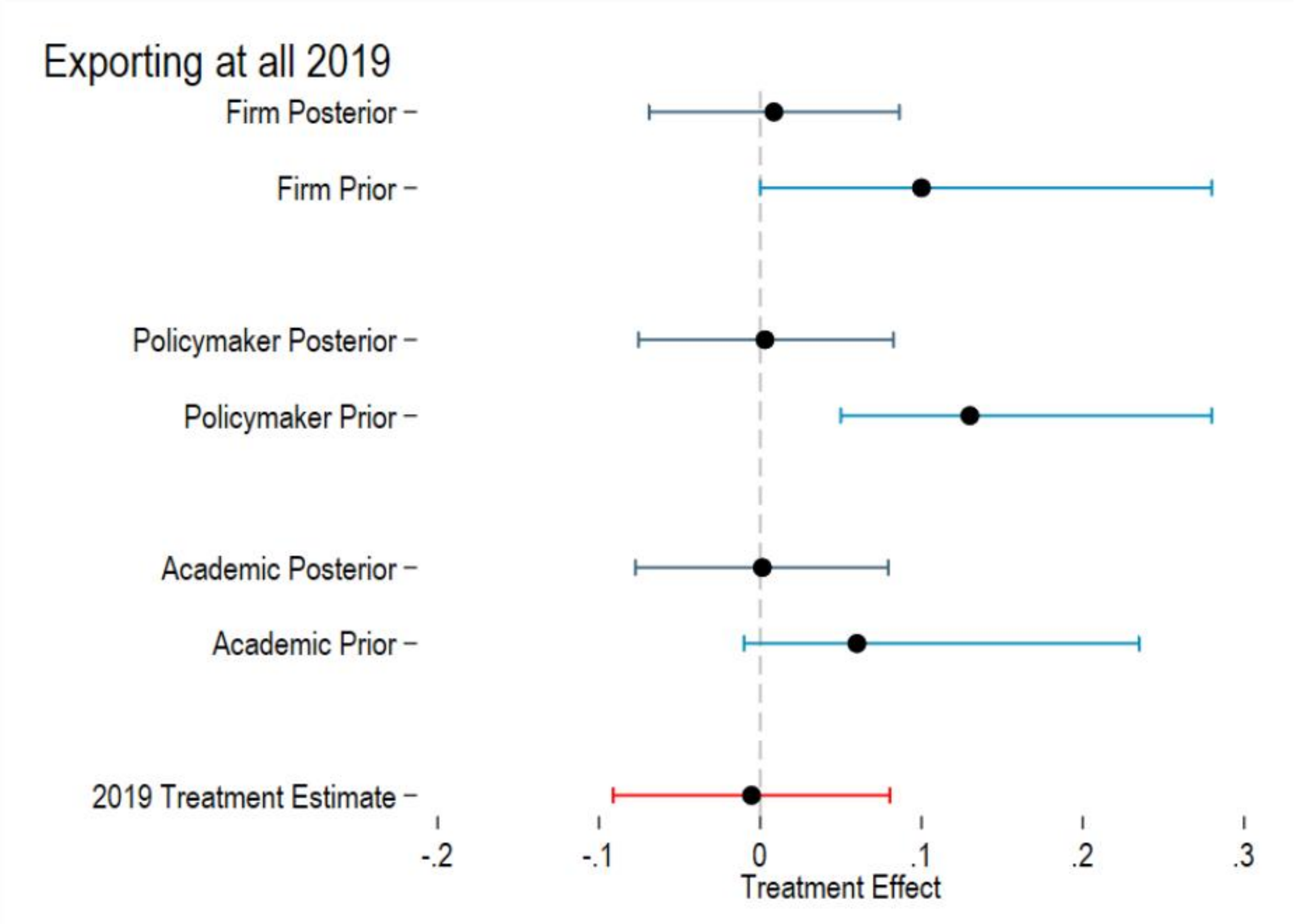
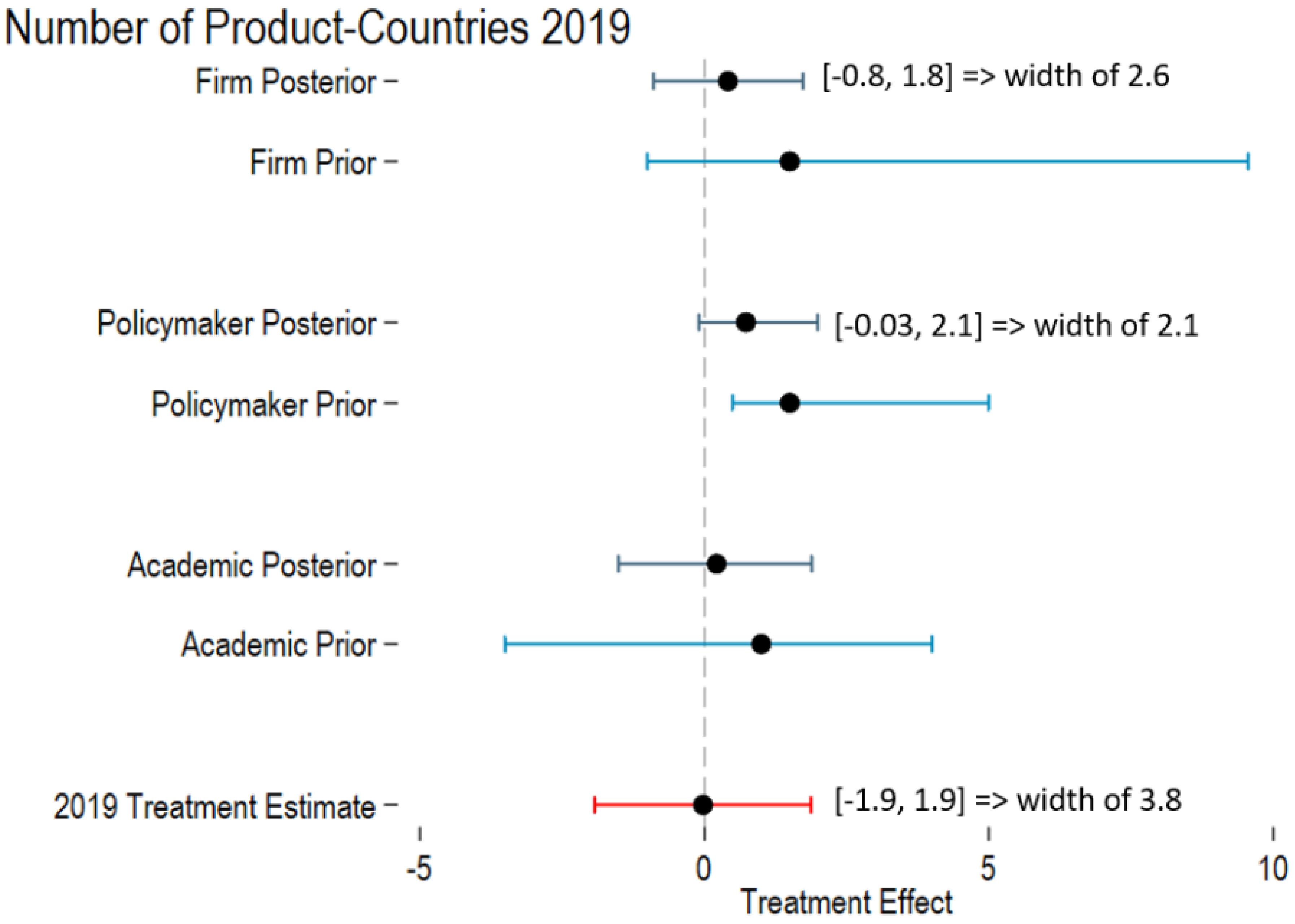


Figure 1: Impact on the Extensive Margin of Whether Firms Export at all



When the data are very informative, priors get almost fully updated and our estimate of the program's effect is what the data shows.

Figure 2: Impact on Export Variety in 2019



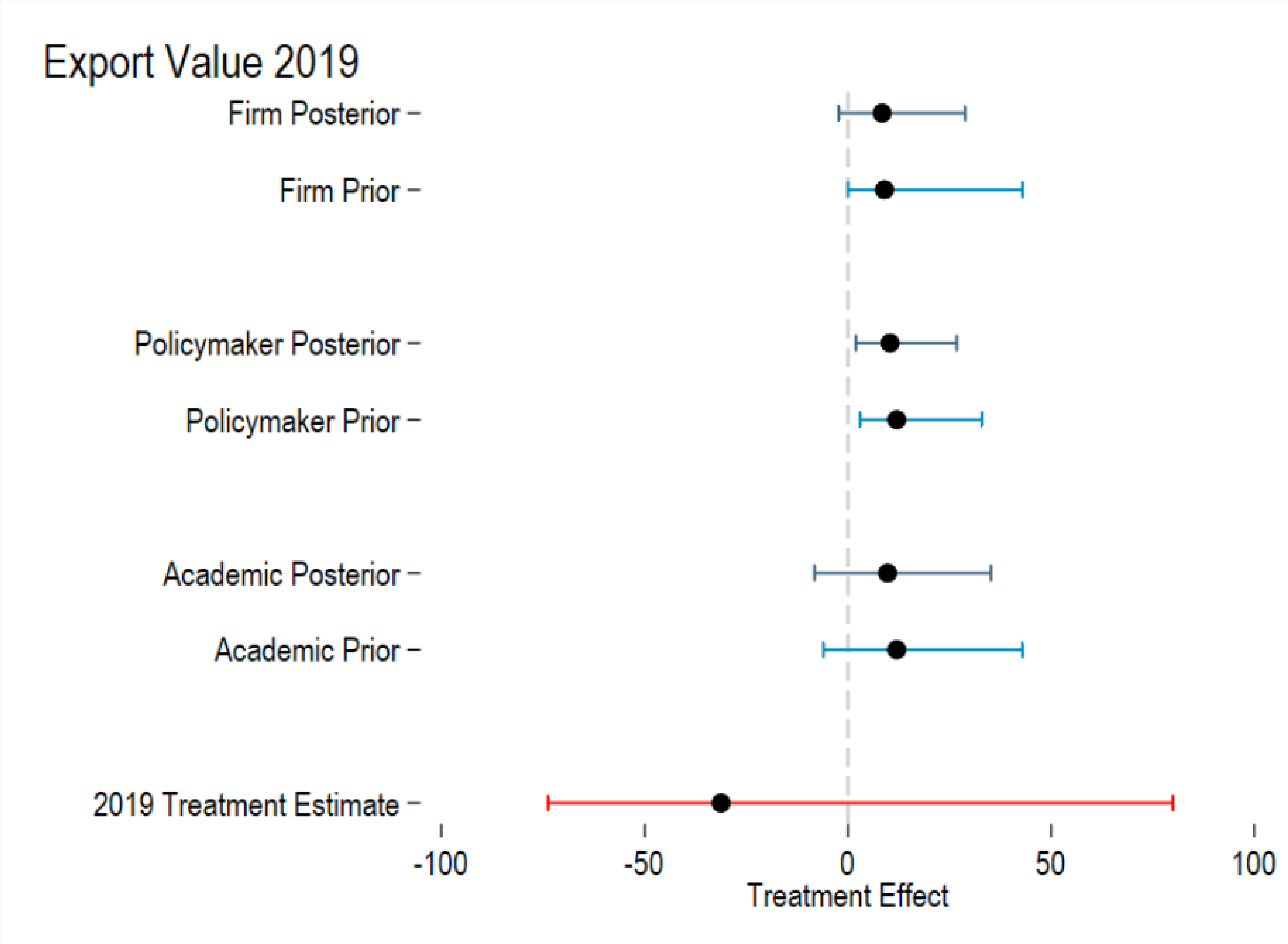
[-0.8, 1.8] => width of 2.6

[-0.03, 2.1] => width of 2.1

[-1.9, 1.9] => width of 3.8

If the data are reasonably informative and in line with our priors, then our posterior intervals can be narrower than with standard analysis

Figure 3: Impact on Export Value



And for some outcomes, we don’t learn much from the experiment, and this analysis tells policymakers they should not update their priors very much.

Example of Bayesian decision analysis

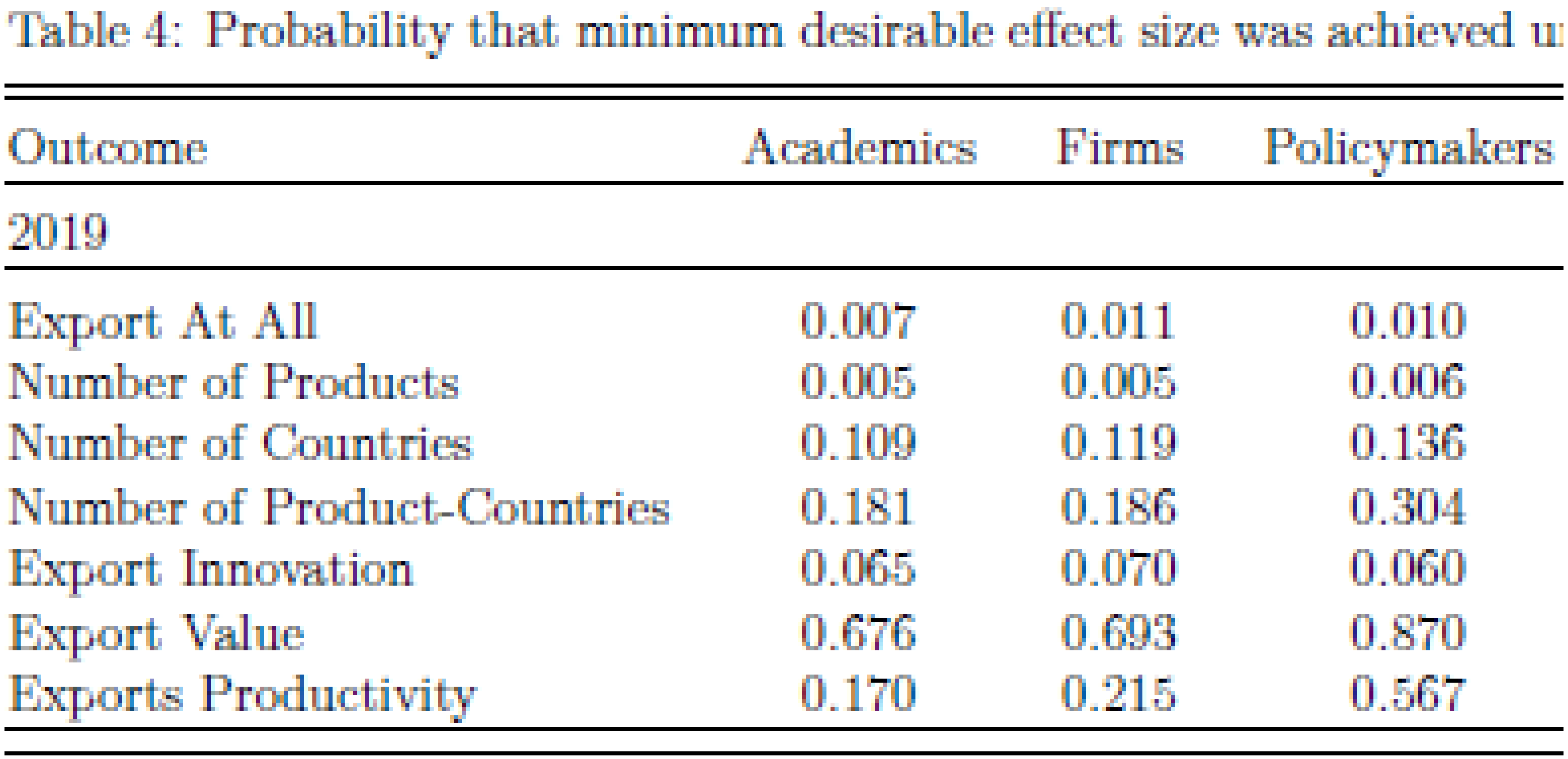


Table 4: Probability that minimum desirable effect size was achieved u

Outcome	Academics	Firms	Policymakers
2019			
Export At All	0.007	0.011	0.010
Number of Products	0.005	0.005	0.006
Number of Countries	0.109	0.119	0.136
Number of Product-Countries	0.181	0.186	0.304
Export Innovation	0.065	0.070	0.060
Export Value	0.676	0.693	0.870
Exports Productivity	0.170	0.215	0.567

When is this approach most useful?

- When sample sizes are small, and it is costly/time-consuming to repeat the experiment
- Other areas where power is limited
  - Multiple treatment arms
  - Treatment heterogeneity

- Provides a way of making sure project is measuring what policymakers want, incorporating their prior information, and helping them update