Discussing causal inference and impact evaluation of Takaful and Karama

by Scott Cunningham (Baylor University)

October 24, 2023

Roadmap

What is causal inference?

- Core questions in causal inference
- Treatment Assignment Mechanisms

Takaful and Karama Impact Evaluation

- Description of program and methodology
- Authors' findings
- Comments and suggestions

Overview of Today's Talk

- Firstly, we'll explore *causal inference*:
  - → What is it? Why is it important? What is at stake?
  - → Importance of *controlled randomization* and options when we can't.
- Secondly, we'll delve into a recent evaluation by Breisinger and coauthors:
  - → Focusing on the Takaful (Solidarity) and Karama (Dignity) programs.
  - → Understanding the evaluation's findings and implications.

Important Distinctions: Causality vs. Causal Inference

What is causal inference? What is causality? They are related but not the same thing.

- *Causality* is a metaphysical concept focusing on the theory of causality
- *Causal Inference* is an epistemological concept focused on how we can know causal effects in data

Important Distinctions: Correlation vs. Causal Inference

What is correlation? When is it causal and when is it not?

- *Correlation* is a purely statistical concept measuring movements between two things
- *Causal Inference* uses in data, assumptions and statistical models to estimate causal effects

Causal Inference Gains

- Causal effects inform policy because they tell us that if we were to undertake a policy, then the estimated causal effects found in our data would likely occur again
- The gains from causal inference help us not only know whether things work (as opposed to driven by spurious and uninformative relationships found in our data sources) but also the magnitude
- Both the direction and the magnitude are critical for policy because they can help determine the size of the impact of what may be expensive programs

Common errors

- Aliens from another planet come and notice that people on ventilators have higher mortality than those not on ventilators
- They conclude that ventilators are killing people
- Are they right? Or they have it backwards – maybe doctors are putting sick people on ventilators to help them
- How can separate the two? By understanding the behaviors that drove people into and out of programs first and combining that with statistical methodologies that take advantage of that

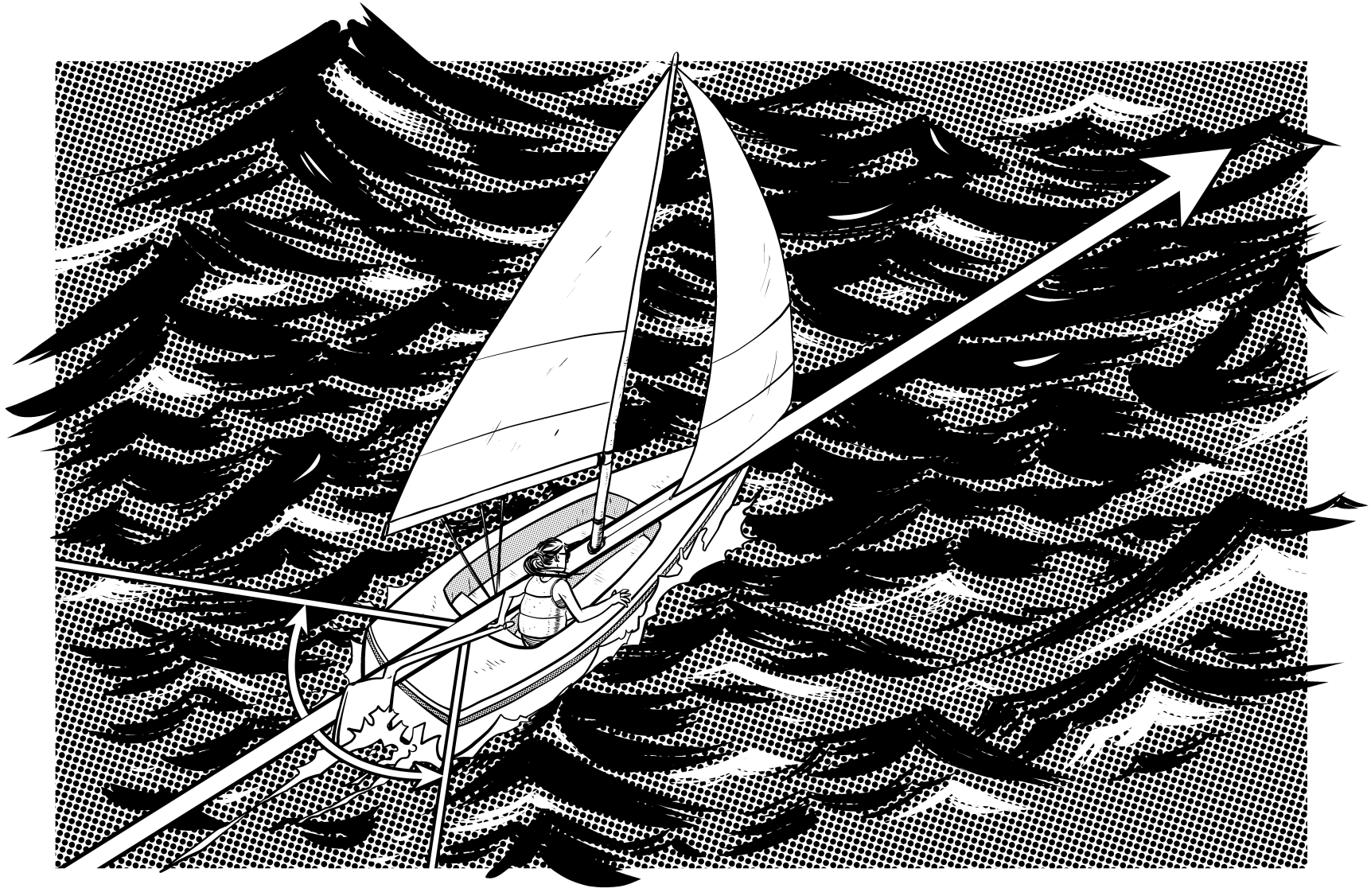
#1: Correlation and causality are different concepts

- Differences between causality and correlation
  - → Causal is about understanding the effect of one unit changing on another. “If a person puts a patient on a ventilator, will her covid symptoms improve?”
  - → Correlation, on the other hand, is about understanding relationships across many units. “How do changes in ventilators relate to changes in covid symptoms across a population?”
- Failure to understand the difference between causal inference and *description* can lead to major errors in assessment and therefore policy recommendations

#2: Coming first may not mean causality!

- Every morning the rooster crows and then the sun rises
- Did the rooster cause the sun to rise? Or did the sun cause the rooster to crow?
- What if cat killed the rooster? Would the sun never rise?
- Simply assuming things happening one after another represents causal effects is an extension of the previous error

#3: Causality may mask correlations!



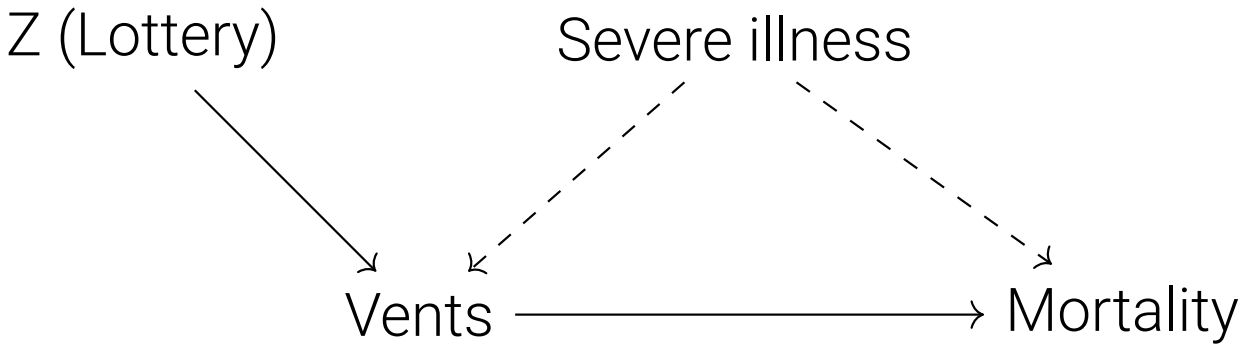
Correlations, Causal Effects, and Selection Bias

- The aliens' error was caused by being unable to separate *selection* from *causal effects*,
    1. **Average causal effect**: What is the average effect of ventilators on mortality?
    2. **Selection bias**: How much of the observed differences between people on and off vents is simply because of baseline differences in mortality that would've been there even had they not been on vents in the first place
- The first one is our goal in policy evaluation, but the second is the problem we seek to overcome with causal inference methods
- Discerning causal effects requires, not statistics, but deciphering the *behavioral reason* that individuals were exposed to some intervention like a poverty program, or what's called the "treatment assignment"

Spectrum of Treatment Assignment

- Treatment assignment *mechanisms* refer to how and why a person participated into a program in the first place and the most common reason – voluntary participation – is a threat to causal inference
- Happens the better run the program is because humans naturally gravitate towards pleasure and away from pain (e.g., vents if COVID symptoms are severe)
- Sometimes the unobserved factors differ between program participants so much that all *observed correlation* between program participants and non-participants is selection bias which leads to incorrect conclusions
- Overcoming this, researchers typically need something that put them into the program other than their own voluntary participation, but what?

Randomization



- Randomized experiments are valuable for causal inference because program participation is *not* based on voluntary participation, and therefore selection bias is minimized almost to nothing
- But for some questions, the *controlled* randomization may not be possible due to feasibility, expense or ethical concerns
- Randomized controlled trials (RCT) use randomization to assign people to vents, but good luck randomizing extremely sick people to ventilators

Running Variables and Regression Discontinuity

- But sometimes people don't choose to participate and aren't even randomized but are assigned to treatment based on a *test*
- When their *grade* on a test is used to put them in a program, we call the test a *running variable* and the eligibility a *cutoff*
- Breisinger and coauthors used this method (also called *regression discontinuity* to study the impact of Takaful and Karama on a variety of health and life outcomes

Roadmap

What is causal inference?

- Core questions in causal inference
- Treatment Assignment Mechanisms

Takaful and Karama Impact Evaluation

- Description of program and methodology
- Authors' findings
- Comments and suggestions

Proxy Means Test

- PMT formula is based on a statistical model measuring log per capita spending and the PMT score used for eligibility into Takaful and Karama was originally 5.003 and lowered to 4.5 for Takaful in Nov 2015
  - → "*PMT [is] an index of well-being based on household demographics, income, housing quality, assets and other characteristics. In poor districts, potentially eligible households were registered and interviewed to collect information for the PMT. Households with a PMT score below a preset threshold were considered eligible for the program and would begin receiving transfers*" – authors
  - → "The PMT has been used to identify the poor within the selected districts, based on selection criteria and a set cutoff score, based on the poverty line derived from Egypt's Household Income, Expenditure and Consumption Survey (HIECS) for 2012/13" – authors
- When this rule is followed perfectly, there is no selection bias when we compare program participants with non-participants, but *only* for the people who just *barely missed* because their score was too low (compared to those who just barely got in)

Changing thresholds

Table 3.2.1 *Takaful* proxy means test score thresholds

Registration period	Dates	Takaful threshold
1	March to November 2015	5003
2	November 2015 to September 2016	4296
3	September 2016 to April 2017	4500
4	April 2017 to present	4500 for male-headed households; 6500 for female-headed households

Table 3.2.2 *Karama* proxy means test score thresholds

Registration period	Dates	Karama threshold
1	March to November 2015	5003
2	November 2015 to May 2016	5063
3	May 2016 to April 2017	7203
4	April 2017 to present	8500

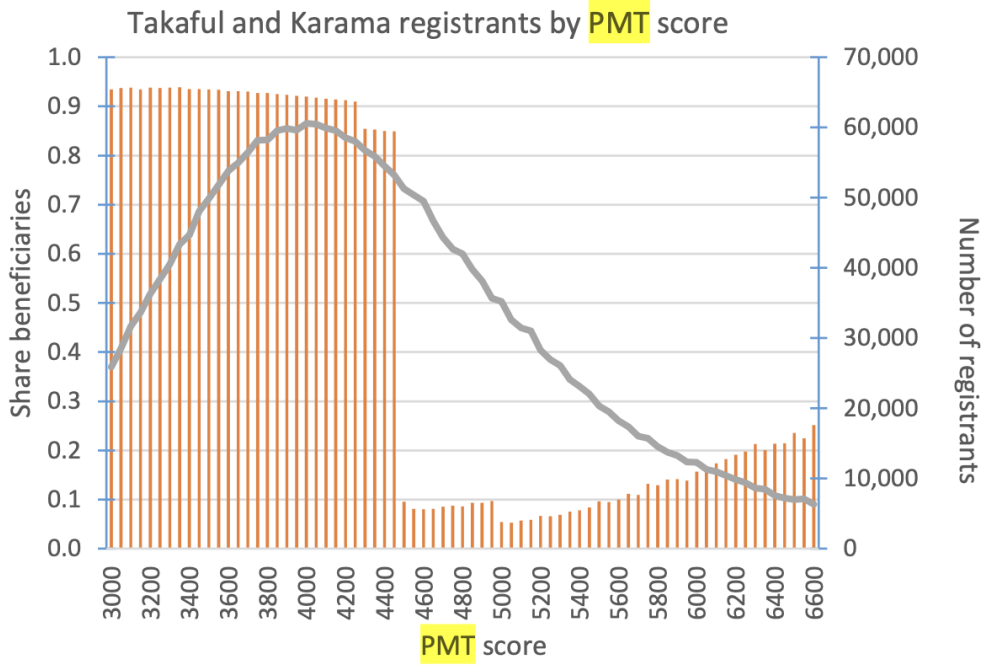
Figure: PMT thresholds over time for both programs

Fuzzy participation

- Does not work, though, if administrators sometimes break the rules (i.e., it isn't followed perfectly)
- Authors note that sometimes people who are eligible still won't participate (sometimes called "non-compliance")
- Authors augment their study to account for this type of *voluntary compliance* using "fuzzy RDD"
- But pretty sharp as you'll see, at least for one of the cutoffs (4.5)

Outcomes

Figure 3.2.1 Beneficiary status in the proxy means test score distribution

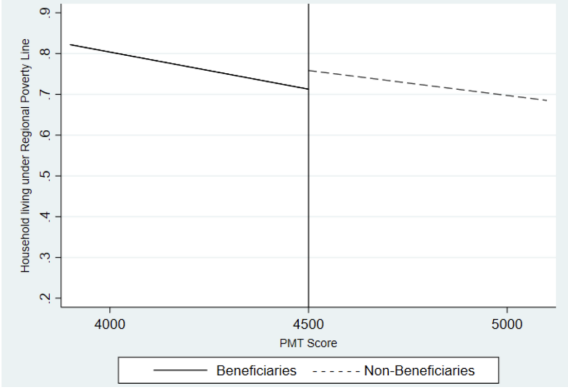


Source: Administrative data from MoSS, received June 2017. Includes only registrants up to April 2017 due to time required to update the database after receiving registration forms.

Figure: Participation by score

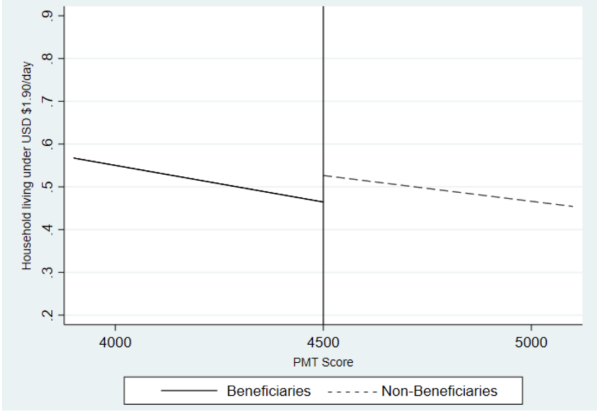
Poverty (top), food and total spending (bottom)

Figure 6.1.4 Regression discontinuity model impact estimates of *Takaful* program on poverty (regional poverty line)



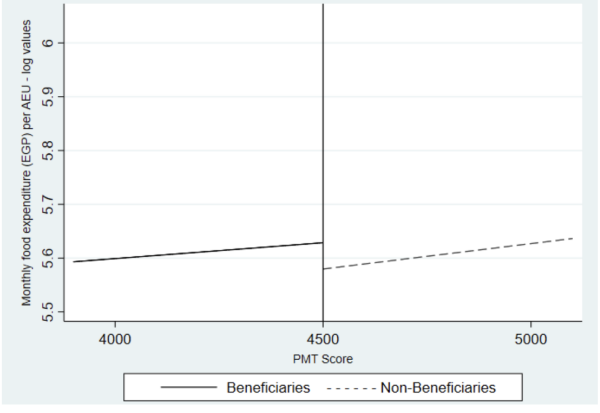
Note: PMT = proxy means test.

Figure 6.1.3 Regression discontinuity model impact estimates of *Takaful* program on poverty (US$1.90/day)



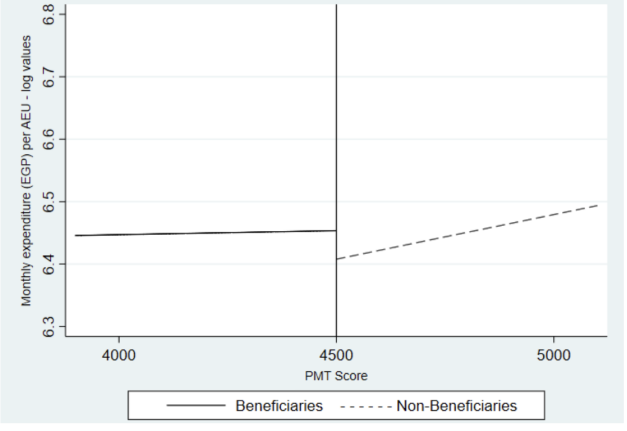
Note: PMT = proxy means test.

Figure 6.1.2 Regression discontinuity model impact estimates of *Takaful* program on log monthly food expenditure per adult equivalent unit



Note: AEU = adult equivalent unit; EGP = Egyptian pounds; PMT = proxy means test.

Figure 6.1.1 Regression discontinuity model impact estimate of the *Takaful* program on log monthly expenditure per adult equivalent unit



Note: AEU = adult equivalent unit; EGP = Egyptian pounds; PMT = proxy means test.

Summary

- Some effects are clearer than others
  - → Large effects on food spending (around 8%) and total spending (around 9%)
  - → Large reductions in poverty (around 12% reduction in living under poverty line)
  - → Huge effects on child “weight-for-length/height” (30-40% SD) and reduced malnourishment (3-4%)
- Some effects, particularly on food and clothes spending are unclear
  - → Increased consumption of fruit (around 25%) but this is a little noisy and only shows up for one model
  - → Large increases in meat consumption (around 30-40%)
  - → Some evidence for increased spending on clothes but also noisy
- Inconsistent evidence for optimism about future, spending on schooling, but some paradoxes like weakened female bargaining power over children schooling and healthcare

Comments

- Very thorough, very contemporary in many ways, very interesting, very valuable – highly encourage people to study it carefully, and re-evaluations done to confirm, as many good news and somewhat bad news (lots of null results)
- RDD and the fuzzy method helps paint a picture that particularly around 4.5 there seems to be some improvements due to the program
- Some things are strange too – like worsened female bargaining power around child welfare, which I think makes this a somewhat intriguing finding meriting more research later
- Much stronger evidence for Takaful than Karama, which is also puzzling

Comments

- Authors estimate the average causal effect of the program *at the thresholds only* and this is a strength and limitation
- Means we can only learn average effects at the cutoff, but if there is large sweeping differences in returns elsewhere, then this will not be informative about program's average effect
- Means the finding has *internal validity* (i.e., they found an average effect) but it lacks *external validity* unless everyone has that same effect
- Future work will need to try and determine to what degree that is the case

More pictures needed

- To really communicate the findings, I think the authors need to find a way to communicate these results without so many tables
- They may want to consider presenting the results using a boxplot of coefficients after normalizing scores into z-scores everywhere (see the following example)
- This could help summarize the results as 150 pages is being chewed up by tons and tons of tables
- RDD plots are hard to evaluate without scatter plot means along the bandwidth – is this noise? Is this linear like they show?

Example of z-score box plots from another paper

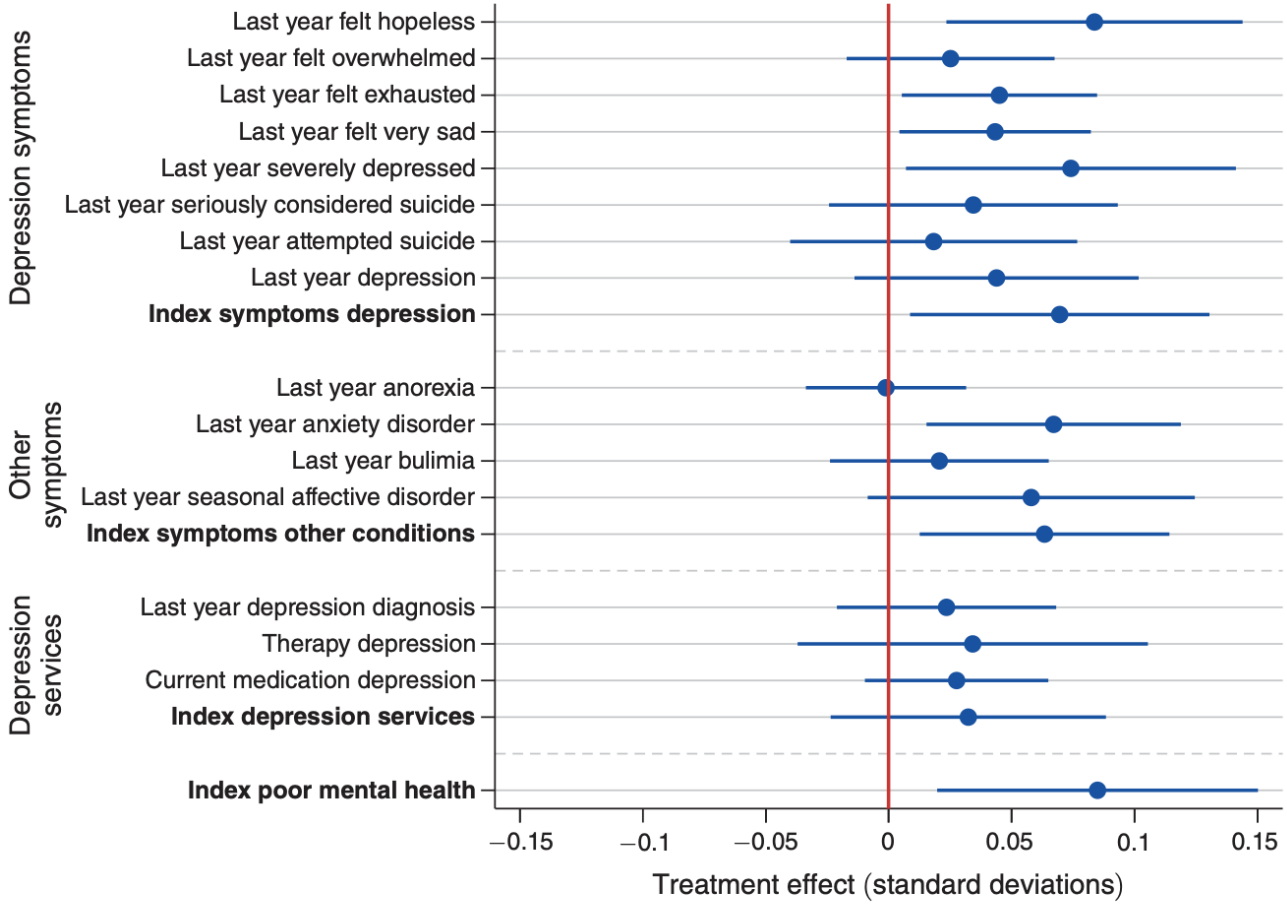


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

Figure: Example of alternative data visualization

Comments

- Authors model uses each threshold, I would prefer to see a single eligibility threshold with a re-centered running PMT score so they can interact eligibility with score
- Doing so will give a *single instrument* which will minimize biases due to weak instruments (and it seems like they're acknowledging that they may have some weak response to the threshold except for 4.5 on PMT)
- Standard first stage statistical tests on the strength of the participation will be needed but looking at the earlier table, it appears particularly strong for 4.5 – just check anyway
- Not sure the first stage measurements of strength are correct if they're using multiple instruments (i.e., multiple cutoffs), but highly recommend using only 4.5 or combining them so you can use the more appropriate "strength of eligibility tests"

Comments

- First and second stage of instrumental variables should have same controls, but authors only control for strata in second stage according to their equation
- Highly encourage the authors to use IV models like `ivregress 2sls` in Stata (or equivalent in R) so that this is guaranteed
- Consider controlling for score and a quadratic in score to model nonlinearities because ultimately RDD “extrapolates” based on the functional form and they may have underfitting

Conclusion

- Paper's strengths are its rigor and focus on strong methods for overcoming selection bias
- Some findings stronger than others
- Causal inference methodologies have advanced in the non-randomized setting, but have not replaced the controlled randomization and never will
- Policymakers can consider studying programs using these methods, and should, but keep in mind limitations and challenges in inference