

About Causal Inference and How It Can Improve Impact Evaluation of Takaful and Karama

by Scott Cunningham (Baylor University)

October 25, 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 2018 evaluation by IFPRI:
 - Focusing on the Takaful (Solidarity) and Karama (Dignity) programs.
 - Understanding the evaluation's findings and implications.

Advances in causal inference

- Causal inference is having its day in the sun
- Explosion in advances in causal inference recently awarded with several major awards:
 - Josh Angrist, David Card and Guido Imbens (2021 Nobel Prize in Economics)
 - Judea Pearl (2011 Turing Award in computer science)
 - James Robins, Miguel Hernán, Thomas Richardson, Andrea Rotnitzky, and Eric Tchetgen Tchetgen (2022 Rousseeuw Prize for Statistics)
- Widespread adoption of causal inference for data driven decision making, both in government and commerce, has replaced simpler correlational methods
- Exciting time!

Examples of causal decisions

Causal inference helps us know whether things work and by how much

- Will mandatory vaccination laws reduce the spread of COVID?
- How will a new early reading curriculum impact youth literacy?
- How much, if any, will cash transfers impact poor family's overall well-being?

Counterfactuals and causal inference

Ladder of causation: observe relationships, intervene, and reason out the counterfactual

Fundamental problem of causal inference: We don't have anyone's counterfactuals because *by definition* they never happened

Causal inference methods *estimate* counterfactuals to understand intervention's impact but as counterfactuals don't exist, estimation require *assumptions*, data and appropriate methods

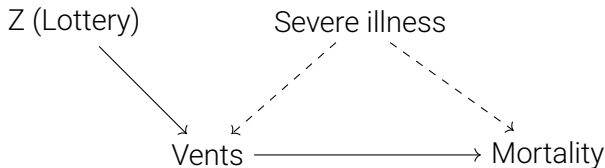
Correlations, Causal Effects, and Selection Bias

- Imagine we find patients on ventilators have higher mortality than patients who aren't; two problems we face
 1. **Average causal effect:** Can we find the average effect of ventilators on mortality? If so how?
 2. **Selection bias:** How much of the observed differences between people on and off vents is because these the ventilator group always would've had higher mortality?
- Our goal: find situations where we can credibly delete the second term so all that is left is the first term when comparing program participants to non-participants
- Requires understanding the *behavioral reason* people were selected, but some behavioral reasons for program participation make this a very difficult problem disentangle

Spectrum of Treatment Assignment

- Ironically, the better run a program is, the more difficult it is to infer the effect
- Voluntary participation in well run programs become dominated by selection bias and may make it impossible to know average causal effects
- Overcoming selection bias requires something external 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
- Randomized assignment of rural regions to good and bad schools would definitively measure the impact of schools on children, but this might be too expensive, infeasible or unethical to knowingly deprive some areas of crucial educational improvements

Running Variables and Regression Discontinuity

- But sometimes people were limited in their voluntary participation, not because they were in a controlled randomized experiment, but because their participation was based on a *test* and their *score*
- 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*
- Impact study by IFPRI published in 2018 used this method (also called *regression discontinuity*) to study the impact of Takaful and Karama on a variety of health and life outcomes

Methodology commentary

- RDD is a design can under certain assumptions identify the average effect of the program on well being but *only* for those people right above and below the eligibility cutoff
- Insofar as those people differ from everyone else, the causal effects may lack external validity despite being true for those people
- Method has a variety of techniques but at their core, they compare program participants who just barely got in because of some score they received on a running variable to those who just barely missed the cutoff
- No more “scientific” or valid than any other method so long as the behavioral assignments meet the assumptions required for inference

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 household characteristics 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 (but raised for Karama)
 - *"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 |

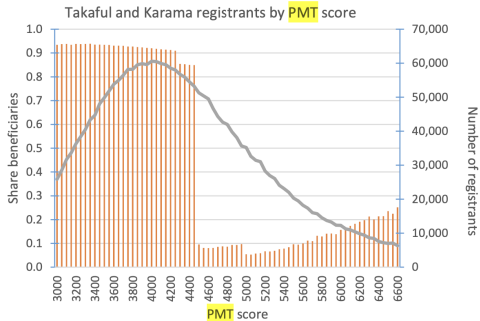
Figure: PMT thresholds over time for both programs

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)

Program Participation at 4.5 PMT

Figure 3.2.1 Beneficiary status in the proxy means test score distribution

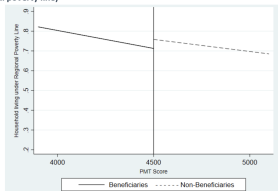


Source: Administrative data from MoSS, received June 2017. Includes only registrants up to April 2017 due to time required to update the database after receiving registration forms.

Figure: Participation by score

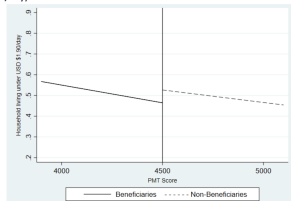
Poverty (top), food and total spending (bottom)

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



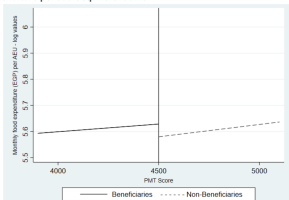
Note: PMT = proxy means test.

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



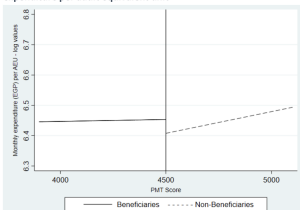
Note: PMT = proxy means test.

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, very interesting, very valuable – highly encourage people to study it carefully, and re-evaluations done to confirm, as many good news (but lots of null results too)
- RDD with and without fuzzy design paint a picture that around 4.5 there are some improvements from Takaful participation
- But 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

Local impact is for some; average impact for everyone

- Authors estimate a “local” average causal effect of the program *at the thresholds only* and this is a strength and limitation
- We can only learn impact for people who just barely missed and barely got in at 4.5, as opposed to all program participants, and policy choices should be based on everyone
- If there is large differences in program effects, then this will not be informative about program’s average effect (external validity), and policy should be judged by everyone, not just a thin slice at some arbitrary cutoff
- Future work will need to try evaluate effects on everyone to determine to what degree that is the case which will require moving away from RDD

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

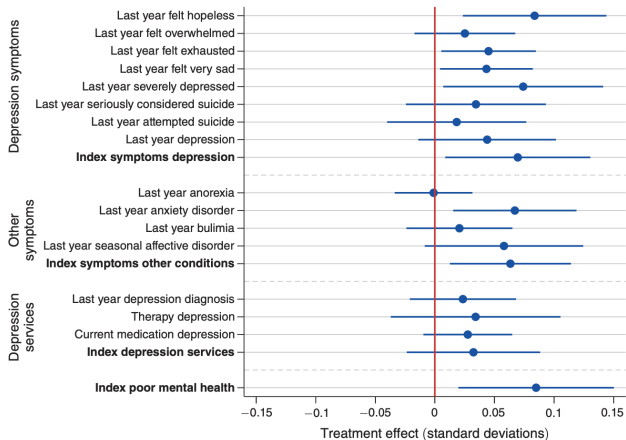


FIGURE 1. EFFECTS OF THE INTRODUCTION OF FACEBOOK ON STUDENT MENTAL HEALTH

Figure: Example of alternative data visualization

Comments on fuzzy model

- Strongest instrument seems to be the 4.5 threshold, but authors are using multiple thresholds
- Would prefer to see a single eligibility threshold with a re-centered running PMT score so they can interact eligibility with score, as this is just identified which makes first stage tests easier and weakness less a problem
- Doing so will give a *single instrument* which will minimize biases due to weak instruments (unclear strength of first stage beyond 4.5 PMT)
- Another reason is it is inappropriate to use either the conventional or heteroskedasticity robust F-test for evaluating first stage strength

Comments on fuzzy model

- 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

Strengths and implications

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

Credible Causal Inference and Parting Remarks

- Validity in causal inference is based on methods that reflect the reality of why people entered a program (i.e., all methods are equally scientific)
- Fuzzy RDD *only* estimates the “local average treatment effect” which is only the average effect for people at 4.5 (internal but not external validity unless effects are the same for everyone)
- Future avenues need to examine the average effect for *all participants* as treatment effect heterogeneity could strengthen or weaken results found at 4.5 participants
- Machine learning methods, such as causal forests and double debiased machine learning, can do this, but require a lot of information on observable confounders related to participation and outcomes
- Thank you!!