MANCHESTER
1824
The University of Manchester

The Use of Experimental and Quasi-Experimental Methods in Innovation Policy Evaluation

Abdullah Gök
Manchester Institute of Innovation Research, University of Manchester, UK

14/11/2013, Vienna

FTEVAL Conference on Evaluation of STI policies, instruments and organisations: new horizons and new challenges

Introduction

- Evaluation Problem: Net Impact of a Policy = Observed Outcome – Unobserved Counterfactual (what would have happened without the policy)
- Evaluation Designs to overcome the evaluation problem
  - Experimental designs (i.e. randomisation or RCTs): randomly selecting a treatment and non-treatment group
  - Quasi-Experimental designs: actively balancing non-random treatment and non-treatment group or using only the treatment group before-after comparison
  - Non-experimental designs: not using a non-treatment group or before-after comparison
- Experimental designs as "gold standards" and hierarchies of evidence
- Experimental designs are rare but quasi-experimental designs are considerable (around 50%)
- Call for more use of experimental and quasi experimental methods in all policy areas

Research Questions:

- How applicable are the experimental and quasi-experimental designs to innovation policy evaluation?
- Are the experimental and quasi experimental designs perceived as more useful and of higher quality by policy-makers?

Method and Structure

- Data Sources:
  - Compendium of Evidence on the Effectiveness of Innovation Policy
    - Evaluation synthesis of 197 evaluation reports and 584 academic publications with evaluation evidence
  - INNO-Appraisal Innovation Policy Evaluations Repository (IPER):
    - Meta evaluation of 171 EU28 innovation policy evaluations (2002-2007) in terms of their questions, methods, topics and audiences
    - Issues of quality and usefulness were assessed by respective policy-makers (N: 132)
- Structure:
  - Innovation policy versus other policy areas
  - Threats to validity and their relevance to innovation policy
    - Statistical Conclusion Validity
    - Internal Validity
    - Construct Validity
    - External Validity
  - Quasi Experimental Designs versus
    - Other evaluation characteristics
    - Perceived quality
    - Perceived Usefulness
  - Conclusion

Innovation Policy versus Other Policy Areas (Or What is Special about Innovation Policy?)

A list of issues that might arise in the evaluation of innovation policy

- ***Paucity:*** Number of units are comparatively very low
- ***Heterogeneity*****:** Units are very heterogeneous (in terms of size, motivation, location, activities, processes etc.)
- ***Fluidity*****:** Units are changing very rapidly and frequently
- ***Long-tailed Effects*****:** Generally very few units have radical effects
- ***Duration:*** Intervention generally spans longer time-frames
- ***Lagged Effects:*** Effects generally occur with a lag
- ***Non-Aggregatability:*** There is not a clear aggregation between different levels of units especially due to evolutionary processes
- ***Low Observability*****:** Making observation is more difficult and often through proxy indicators
- ***Complex Policy:*** Policy logic is more complex:
  - Innovation policy aims to encourage units to do something differently
  - Ultimate objective of innovation policy is difficult to measure
  - Often there are intermediate targets
  - Often regular but unexpected punctuated equilibria
- ***Complex mix of effects:*** There is a complex interplay of a variety of effects in influencing what innovation policy targets
- ***Endogeneity***: Endogeneity is much more common especially due to the cause-effect loop
- **Strategic Behaviour:** Units respond strategically to the policy and evaluation and change their position rapidly after

Threats to Validity

Shadish et al. 2002 framework

- **Statistical Conclusion Validity:** How valid is the statistical inference between the cause and effect?
  - Low Statistical Power *(due to chronic paucity)*
  - Violated Assumptions of Statistical Tests *(due to heterogeneity, long-tailed effects)*
  - Heterogeneity of Units
- **Internal Validity:** How valid is the inference of the relationship between the cause and effect?
  - Ambiguous Temporal Precedence *(due to endogeneity, lagged effects, duration)*
  - Selection *(due to heterogeneity, complex policy)*
  - History *(due to complex mix of effects, fluidity)*
  - Maturation *(due to complex mix of effects, fluidity)*
- **External Validity:** How valid is the generalisation of the inference to other circumstances?
  - Interaction of the Causal Relationship with Units *(due to heterogeneity)*
  - Interaction of the Causal Relationship Over Treatment Variations *(due to complex policy, complex mix of effects, heterogeneity)*
  - Interaction of the Causal Relationship with Outcomes *(due to complex policy, complex mix of effects, heterogeneity)*
  - Interactions of the Causal Relationship with Settings *(due to complex mix of effects, fluidity, non-aggregatability)*
- **Construct Validity:** How valid is the operationalisation of the evaluation in evaluating the relationship between the cause and effect?
  - Inadequate Explication of Constructs *(due to complex policy)*
  - Reactive Self-Report Changes *(due to strategic behaviour)*

Threats to Statistical Conclusion Validity

Threat	Issue / Definition[1]	Relevance to Innovation Policy
Low Statistical Power	An insufficiently powered experiment may incorrectly conclude that the relationship between treatment and outcome is not significant.	+++ *chronic paucity*
Violated Assumptions of Statistical Tests	Violations of statistical test assumptions can lead to either overestimating or underestimating the size and significance of an effect.	+++ *heterogeneity, long-tailed effects*
Fishing and the Error Rate Problem	Repeated tests for significant relationships, if uncorrected for the number of tests, can artifactually inflate statistical significance.	+ *low observability*
Unreliability of Measures	Measurement error weakens the relationship between two variables and strengthens or weakens the relationships among three or more variables.	+ *low observability*
Restriction of Range	Reduced range on a variable usually weakens the relationship between it and another variable.	+ *low observability*
Unreliability of Treatment Implementation	If a treatment that is intended to be implemented in a standardized manner is implemented only partially for some respondents, effects may be underestimated compared with full implementation.	+ *complex policy*
Extraneous Variance in the Experimental Setting	Some features of an experimental setting may inflate error, making detection of an effect more difficult.	
Heterogeneity of Units	Increased variability on the outcome variable within conditions increases error variance, making detection of a relationship more difficult.	+++ *heterogeneity*
Inaccurate Effect Size Estimation	Some statistics systematically overestimate or underestimate the size of an effect.	

[1] Based on Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

The University of Manchester

Threats to Internal Validity

Threat	Issue / Definition[1]	Relevance to Innovation Policy
Ambiguous Temporal Precedence	Lack of clarity about which variable occurred first may yield confusion about which variable is the cause and which is the effect.	++ *Endogeneity, lagged effects, duration*
Selection	Systematic differences over conditions in respondent characteristics that could also cause the observed effect.	+++ *Heterogeneity, complex policy*
History	Events occurring concurrently with treatment could cause the observed effect.	++ *Complex mix of effects, Fluidity*
Maturation	Naturally occurring changes over time could be confused with a treatment effect.	++ *Complex mix of effects, Fluidity*
Regression	When units are selected for their extreme scores, they will often have less extreme scores on other variables, an occurrence that can be confused with a treatment effect.	
Attrition	Loss of respondents to treatment or to measurement can produce artifactual effects if that loss is systematically correlated with conditions.	+ *Fluidity*
Testing	Exposure to a test can affect scores on subsequent exposures to that test, an occurrence that can be confused with a treatment effect.	+ *Strategic Behaviour*
Instrumentation	The nature of a measure may change over time or conditions in a way that could be confused with a treatment effect.	+ *complex policy*
Additive and Interactive Effects of Threats to Internal Validity	The impact of a threat can be added to that of another threat or may depend on the level of another threat.	

[1] Based on Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Threats to External Validity

Threat	Issue / Definition[1]	Relevance to Innovation Policy
Interaction of the Causal Relationship with Units	Certain kinds of units might not hold if other kinds of units had been studied.	++ *Heterogeneity*
Interaction of the Causal Relationship Over Treatment Variations	One treatment variation might not hold with other variations of that treatment, or when that treatment is combined with other treatments, or when only part of that treatment is used.	+++ *Complex policy, Complex mix of effects, Heterogeneity*
Interaction of the Causal Relationship with Outcomes	One kind of outcome observation may not hold if other outcome observations were used.	++ *Complex policy, Complex mix of effects, Heterogeneity*
Interactions of the Causal Relationship with Settings	One kind of setting may not hold if other kinds of settings were to be used.	++ *Complex mix of effects, Fluidity, Non-Aggregatability*
Context-Dependent Mediation	An explanatory mediator of a causal relationship in one context may not mediate in another context.	+ *Complex policy*

[1] Based on Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

The University of Manchester

Threats to Construct Validity

Threat	Issue / Definition[1]	Relevance to Innovation Policy
Inadequate Explication of Constructs	Failure to adequately explicate a construct may lead to incorrect inferences about the relationship between operation and construct.	++ *Complex policy*
Construct Confounding	Operations usually involve more than one construct, and failure to describe all the constructs may result in incomplete construct inferences.	
Mono-Operation Bias	Any one operationalization of a construct both underrepresents the construct of interest and measures irrelevant constructs, complicating inference.	+ *low observability*
Mono-Method Bias	When all operationalizations use the same method (e.g., self-report), this method is part of the construct actually studied.	+ *low observability*
Confounding Constructs with Levels of Constructs	Inferences about the constructs that best represent study operations may fail to describe the limited levels of the construct that were actually studied.	
Treatment Sensitive Factorial Structure	The structure of a measure may change as a result of treatment, change that may be hidden if the same scoring is always used.	
Reactive Self-Report Changes	Self-reports can be affected by participant motivation to be in a treatment condition, motivation that can change after assignment is made.	++ *strategic behaviour*
Reactivity to the Experimental Situation	Participant responses reflect not just treatments and measures but also participants' perceptions of the experimental situation, and those perceptions are part of the treatment construct actually tested.	++ *strategic behaviour*

[1] Based on Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Threats to Construct Validity

Threat	Issue / Definition[1]	Relevance to Innovation Policy
Experimenter Expectancies	The experimenter can influence participant responses by conveying expectations about desirable responses, and those expectations are part of the treatment construct as actually tested.	
Novelty and Disruption Effects	Participants may respond unusually well to a novel innovation or unusually poorly to one that disrupts their routine, a response that must then be included as part of the treatment construct description.	
Compensatory Equalization	When treatment provides desirable goods or services, administrators, staff, or constituents may provide compensatory goods or services to those not receiving treatment, and this action must then be included as part of the treatment construct description.	
Compensatory Rivalry	Participants not receiving treatment may be motivated to show they can do as well as those receiving treatment, and this compensatory rivalry must then be included as part of the treatment construct description.	
Resentful Demoralization	Participants not receiving a desirable treatment may be so resentful or demoralized that they may respond more negatively than otherwise, and this resentful demoralization must then be included as part of the treatment construct description.	✚ *Complex policy*
Treatment Diffusion	Participants may receive services from a condition to which they were not assigned, making construct descriptions of both conditions more difficult.	

[1] Based on Shadish, W.R., Cook, T.D., Campbell, D.T., 2002. Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.

Quasi Experimental Methods and Other Evaluation Characteristics

- Based on the analysis of INNO-Appraisal Database (N=171 evaluations), statistically significant associations (Chi-Square test):
- Control Group
  - More: Summative, concerned with input and output additionality (but not behavioural additionality), economic (rather than scientific, technological, social and environmental) impact, econometric methods, policy analyst audience
  - Less: qualitative methods (interviews, case studies, etc.)
- Before/After Comparison
  - More: *Summative+Formative mixed*, concerned with input and output additionality (but not behavioural additionality), gender and minority issues, economic (rather than scientific, technological, social and environmental) impact, econometric methods, policy analyst audience
  - Less: qualitative methods (interviews, case studies, etc.)

MANCHESTER 1824 The University of Manchester

Quasi Experimental Methods and Perceived Quality

- Usefulness of recommendation defined in 5 point Likert Scale (1-Not at all to 5- Extensive)
- Overall N=132, verified by respective policy makers
- Independent Samples Means t-test, 2 tailed, variance equality also tested and adjusted

Perceived Quality Dimension	Status of Column Variable	Control Group Approach			Before/After Group Comparison Approach		
		N	Mean	Sig (2 tailed)	N	Mean	Sig (2 tailed)
Was/Is the design of the evaluation appropriate given the objectives of the evaluation and the nature of the policy measure?	No	89	4.01	.264	101	4.02	.307
	Yes	21	4.24		10	4.30	
Did/Do the methods chosen satisfy the Terms of Reference/purpose of the appraisal?	No	74	4.22	.132	84	4.25	.238
	Yes	18	4.56		8	4.63	
Was/Is the application of the qualitative methods satisfactory?	No	85	3.95	.801	95	3.89	.161
	Yes	19	3.89		9	4.33	
Was/Is the application of the quantitative methods satisfactory?	No	**78**	**3.67**	**.001**	89	3.80	.553
	Yes	**20**	**4.45**		9	4.00	
Were/Are the information sources used in the report well documented and referenced?	No	89	4.18	.734	102	4.15	.884
	Yes	22	4.09		10	4.20	
Was/Is the analysis clearly based on the data given?	No	89	4.22	.125	100	4.24	.257
	Yes	22	4.50		12	4.50	
Given the objectives of the appraisal, does the analysis cover the broader context (e.g. societal, institutional, policy and economic contexts) sufficiently?	No	89	3.36	.168	**98**	**3.34**	**.016**
	Yes	20	3.75		**12**	**4.17**	
Were/Are the conclusions based on the analysis?	No	90	4.29	.149	**101**	**4.30**	**.090**
	Yes	22	4.59		**12**	**4.75**	

Quasi Experimental Methods and Perceived Usefulness

- Usefulness of recommendation defined in 5 point Likert Scale (1-Not at all to 5- Extensive)
- Overall N=132, verified by respective policy makers
- Independent Samples Means t-test, 2 tailed, variance equality also tested and adjusted

Perceived Usefulness Dimension	Status of Column Variable	Control Group Approach			Before/After Group Comparison Approach		
		N	Mean	Sig (2 tailed)	N	Mean	Sig (2 tailed)
Changes to the design of the programme/measure appraised	No	65	3.02	.810	63	3.05	.424
	Yes	8	3.13		11	2.73	
Changes to the management and implementation of the programme/measure appraised	No	68	3.28	.571	66	3.32	.238
	Yes	7	3.00		10	2.70	
Changes to the design, management and implementation of future programmes/measures	No	68	3.56	.591	69	3.65	.103
	Yes	9	3.78		9	2.78	
Changes to the design, management and implementation of contemporaneous programmes/measures	No	51	2.18	.257	**55**	**2.35**	**.045**
	Yes	9	2.67		**6**	**1.33**	
Changes to broader policy formulation and implementation	No	62	2.79	.134	**66**	**2.95**	**.029**
	Yes	11	3.36		**8**	**2.13**	

MANCHESTER 1824
The University of Manchester

Conclusions

- Experimental and (in some cases quasi-experimental designs) might generally be less applicable to innovation policy (relative to some other policy areas)
- Although there may be important opportunities where experimental designs can be employed, they are not necessarily the gold standards
- Experimental and Quasi Experimental methods are generally more associated with summative evaluations and economic impacts and econometric analysis
- Quasi-Experimental designs are not perceived as of more quality and useful by policy-makers
- Quality and especially usefulness depend on many other (political) factors
- Design – Quality – Usefulness relationship
  - Appropriate design increases quality to a certain level (but not higher)
  - Good quality increases usefulness to a certain level (but not higher)
- Policy experimentation versus experiments in evaluation

Thank You!

Questions, Comments, Remarks:

abdullah.gok@manchester.ac.uk

“Those are my principles, and if you don't like them... well, I have others!”

Groucho Marx