FORMULATING HYPOTHESES IN THE PROCESS OF DEVELOPING A RESEARCH DESIGN, SELECTING A SYSTEM OF VARIABLES , AND JUSTIFYING AN ECONOMETRIC IDENTIFICATION STRATEGY

Usmonov Maxsud Tulqin o'g'li

ORCID: https://orcid.org/0000-0001-9997-6617 Email: maqsudu32@gmail.com

Qodirov Farrux Ergash o‘g‘li

Scientific advisor: Shahrisabz State Pedagogical Institute Mathematics and information technology in education Head of the Department, ifd DSc, Associate Professor Farrukhbek0209@mail.ru ; orcid.org/0000-0002-4574-7728

Abstract: *This paper discusses the core steps of developing an empirical research design: hypothesis formulation , variable system selection , and justification of an econometric identification strategy . The research workflow is treated as a coherent chain —“ problem–theory–hypothesis–measurement–identification–estimation–validation”—and the paper systematizes the main threats to internal validity (endogeneity, reverse causality, omitted variable bias, selection, measurement error) and practical ways to mitigate them. We show how to convert theoretical statements into testable and falsifiable hypotheses, operationalize constructs, and choose indicators and proxy variables. The paper then reviews major identification strategies—randomization, natural experiments, instrumental variables, difference-in-differences, regression discontinuity, panel fixed effects, matching, and synthetic control—emphasizing their assumptions and diagnostic checks. In the "Results" section, an illustrative scenario of evaluating an education program is presented with explicit hypotheses, a variable matrix, tables, and conceptual figures (a design-selection flowchart and a causal diagram). The paper concludes with actionable guidance on transparency, robustness, and replicability in econometric research.*

Keyword: *research design; hypothesis; variable selection; operationalization; identification; endogeneity; causal inference; instrumental variables; DiD; RD; panel data.*

INTRODUCTION

The value of scientific research often depends on a clear question and a reliable design , rather than on an “interesting result.” Particularly in economics and the social sciences, a robust research design is essential to turn the observation that “the indicators have changed” into a conclusion that “there is a causal effect.” Therefore, in the process of developing a research design, (i) correctly formulating hypotheses, (ii) selecting a system of variables in accordance with the conceptual model, (iii) justifying the identification strategy, and (iv) planning robustness checks play a crucial role [Wooldridge, 2010, 1–10].

The purpose of this article is to provide the practical researcher with a methodical answer to the question “how do I build a design?”: to make the hypothesis testable, operationalize the variables, anticipate the risk of endogeneity, choose an identification path, and defend the choice with empirical evidence. Since issues of causal inference have become a central theme in the contemporary econometric literature, identification strategies are understood as a set of assumptions and design decisions , not just formulas [Angrist and Pischke, 2009, 3–20].

, the concept of hypothesis and its role in design are explained; then, the rules for selecting a system of variables (conceptual model → indicator/proxy → measurement) are given; then, identification strategies, their assumptions and diagnostics are discussed; in the "Results" section, a design example with tables and figures is presented on the example of an educational program; finally, conclusions and recommendations are made.

LITERATURE ANALYSIS AND METHODS

1. Hypothesis: from a theoretical statement to a testable claim

A hypothesis is a statement that presents a theoretical relationship in a format suitable for empirical testing. It should include at least three elements : (1) direction of effect (±), (2) unit of effect (to whom, when), (3) measurable outcome (Y), and cause (X). For example, the general statement "Program coverage increases, academic performance improves" should be transformed into a testable hypothesis such as "Participation in program (X) will increase test scores (Y) on average over 6 months" [Stock and Watson, 2015, 35–48].

In causal research, the hypothesis is usually based on the logic of potential outcomes: for each unit (individual/school/region), the difference in outcomes between the "with " and "without" cases of the program provides a causal effect; however, the incomplete observation of this difference creates an identification problem [Imbens and Rubin, 2015, 11–30]. Therefore, before writing a hypothesis, the questions "what am I comparing?" and "is the comparison fair?" are asked .

Methodological rule 1 (falsifiability): a hypothesis must be falsifiable ; otherwise it fails scientific testing [Popper, 2002, 18–25].

Methodological rule 2 (pre-specified): The risk of p-hacking and "outcome fitting" is reduced if the hypothesis, primary outcome , and underlying model are specified as much in advance as possible [Cameron and Trivedi, 2005, 68–75].

2. System of variables : from conceptual model to operational indicator

of variables is not just a "list of X and Y", but:

- conceptual constructs (e.g., "educational quality", "motivation"),
- their indicators/proxies (test scores, attendance, teacher experience),
- controls (confounders),
- mechanism variables (mediators),
- measures of heterogeneity (subgroup moderators),
- constraints related to data quality (measurement/availability) [Greene, 2012, 80–95].

2.1. "DAG logic" of variable selection

In modern causal analysis, the DAG (Directed Acyclic Graph) is used to explain which variables should be controlled for: the goal is to block out confounders to isolate the effect of X → Y, but not to "over-control" the mechanisms (mediators) and " cut off " the true effect [Pearl, 2009, 17–35]. For example, if a program (X) changes a teacher 's methodology (M) and affects an outcome (Y), controlling for M may measure the "direct effect" rather than the "total effect."

2.2. Operationalization: indicator, proxy, index

Many economic constructs cannot be measured directly. In that case, proxy variables are chosen. Criteria for choosing a proxy:

1. theoretical closeness (construct validity),
2. of measurement error (attenuation bias),
3. possibility of confounding by other factors,
4. stability and repeatability of the information source [Wooldridge, 2010, 300–312].

When constructing indices (e.g., a socioeconomic status index), issues such as standardization, weighting (PCA or expert weighting), and internal reliability (alpha) of the index fall into the "measurement layer" of the design [Gujarati and Porter, 2009, 104–120].

3. Identification: a strategy for "finding" the causal effect

The identification strategy is a framework that specifies "what information and under what assumptions" a causal parameter can be identified . Ordinary OLS is often unbiased due to endogeneity; therefore, the design pre-maps the sources of endogeneity: (i) omitted variables, (ii) reverse causality, (iii) simultaneity, (iv) selection, (v) measurement error [Wooldridge, 2010, 50–65].

Below are the most commonly used identification approaches.

3.1. Randomization (RCT)

In the case of random assignment , X is exogenous and the average causal effect is estimated by simple difference. However, in practice, problems such as attrition, compliance (non-compliance), and spillover occur [Duflo, Glennerster, & Kremer, 2007, 3910–3925]. Diagnostics: balance tests, attrition analysis, ITT/TOT separation.

3.2. Natural experiments and instrumental variables (IV)

The IV approach separates the endogenous part of X through an "external" instrument Z. The main conditions are: relevance (Z shifts X) and exogenousity (ZY affects only through X) [Angrist and Pischke, 2009, 113–140]. Diagnostics: first-stage F-stat, overidentification (if there are many instruments), risk of weak IV [Stock and Watson, 2015, 470–490].

3.3. Difference of Differences (DiD)

measures the "difference in change" between the group that received the policy/program and the group that did not. Basic assumption: parallel trend (in the absence of the intervention , the trends would be parallel) [Angrist and Pischke, 2009, 165–190]. Diagnostics: pre-trend analysis, event-study plots, nested "sham intervention" trials [Abadie, 2005, 5–15].

3.4. Regression Discontinuity (RD)

based on a cutoff (e.g., score ≥ c) , the units around the cutoff are "almost random". The main requirement: no manipulation around the cutoff and careful selection of the functional form [Imbens and Lemieux, 2008, 615–630]. Diagnostics: density test, bandwidth sensitivity.

3.5. Panel data: fixed effects

Time-invariant factors that are not observed in the panel can be controlled with fixed effects . However, time - varying omitted factors, selection, and dynamic endogeneity may remain [Wooldridge, 2010, 280–295]. Diagnostics: cluster-robust SE, trend addition , placebo.

3.6. Matching and synthetic control

in terms of observables ; the main assumption is selection on observables [Rosenbaum and Rubin, 1983, 41–55]. Synthetic control creates an "artificial control" by combining weights for a region; the assumption is that the path of the outcome would converge to this combination in the absence of intervention [Abadie, Diamond, and Hainmueller, 2010, 493–505].

DISCUSSION

1. The biggest design mistake is to ask the wrong question.

Many studies get bogged down in model selection and fail to clarify the question. A causal question should have 4 dimensions:

1. unit (who? — student , class, school),
2. intervention (what? — program, credit, policy),
3. time horizon (when? — 3 months, 1 year),
4. result (what changes ? — score, income, employment).

This clarity makes the hypothesis "operational" and facilitates the selection of a system of variables .

2. The problem of over-incrementing variables

When there are many controls , the intuition that "if I include them all, I will be unbiased" is not always correct. On the one hand, unbiasedness is violated if the necessary confounder is not controlled; on the other hand, bias may appear if the mediator/collider is controlled [Pearl, 2009, 50–65]. Therefore, the selection of controls should be based on a conceptual model (DAG).

How to "justify" an identification strategy ?

The justification consists of three layers:

- Institutional/conceptual framework: why is X exogenous? why is a parallel trend logically expected? is the cutoff not manipulated?
- Empirical diagnostics: balance, pre-trend, density test, first-stage strength.
- Robustness: alternative specification, placebo, sub-sample, heterogeneity.

Many studies do not provide a diagnosis or only provide a single robustness. In fact, the reliability of a strategy is built on the triad of "assumption + evidence + robustness" [Cameron and Trivedi, 2005, 82–90].

4. Ethics, transparency and repeatability

Research design is not only a method, but also a transparent process : the reliability of the results increases if the data source, code, preprocessing, outlier rule, missing data treatment, and model choices are documented. A pre-analysis plan is particularly useful in RCTs and natural experiments [Duflo, Glennerster, & Kremer, 2007, 3925–3935].

RESULTS

we illustrate the methodology in a “sample study” scenario .
Scenario: Effect of additional preparatory course (X) on students' math test (Y)
Question: Does taking a course increase test scores over 1 semester?
Unit: student i, school s, time t.
Intervention: course enrollment/attendance.
Outcome: standardized math score.
Hypotheses

- H1 (main): Attending a course (X) increases test score (Y).
- H2 (mechanism): X → time allocated to homework (M) increases → Y increases.
- H3 (heterogeneity): The effect is stronger for students with low initial scores.

Table 1. Hypothesis–model– measurement fit (design map)

Element	Description	Operational measurement	Expected character	Risk (bias)	Reduction
Y (result)	Mathematics knowledge level	Standardized test score (z-score)	—	measurement error	same test, proctored
X (impact)	Attending the course	0/1 participation; attendance %	+	selection (motivation)	DiD/FE or IV
C (confounder)	Socioeconomic status	SES index	±	omitted variable	Enter SES
C	Basic knowledge	pre-test score	+	reverse causality	pre-test control
M (mechanism)	Independent reading time	weekly hours	+	mediator control risk	the mechanism is in a separate model
Moderator	Low/high pre-test	quantile groups	—	incorrect grouping	predetermination

Table 2. Matrix of variables (minimum + extended specification)

Group	Variable	Marking	Type	Source	Note
Home	Course participation	X_it	binary/continuous	attendance log	ITT vs TOT can be distinguished
Result	Test score	Y_it	continuous	test result	z-score
Basic control	Pre-test	Y_i0	continuous	test	initial differences
Demographic	Age, gender	Age_i, Male_i	control	questionnaire	for heterogeneity
SES	Parental education, income proxy	SES_i	index	questionnaire	PCA/expert
School	Teacher experience	TExp_s	control	administrative	Be careful if it is FE
Time	Semester	τ _t	FE	calendar	general trend

	dummies				

Table 3. Choosing an identification strategy: options and assumptions

Strategy	When is it appropriate?	Basic assumption	Diagnostics	Weak spot
OLS + control	rich data, low endogeneity	Oh, it's	X,C]=0	specification check
Panel FE	there is a time-invariant latent factor	time-invariant u_i disappears	cluster SE, trend	selection by time
DiD	the course was introduced at some point	parallel trend	pre-trend, event study	trend difference
IV	There is a powerful instrument.	relevance+exogenity	first-stage F, placebo	weak IV, exclusion
RD	admission based on cutoff	no manipulation	density test, bandwidth	local effect

CONCLUSION

Building a quality research design is a process that begins before the regression is "written." The design focuses on three issues: (1) making the hypothesis testable and falsifiable ; (2) operationalizing the system of variables based on the conceptual model; and (3) justifying the identification strategy with assumptions, diagnostics, and robustness.

The article put forward the following practical conclusions:

1. Always write the hypothesis in the "who–when–what outcome" format; define the main outcome indicator in advance .
2. Variable selection should be based on DAG logic: block confounders, be careful with mediators/colliders.
3. When choosing an identification strategy, it is not the "strongest method", but the "most appropriate design for the context" that prevails: the sequence RCT → RD → DiD → FE → IV → matching/synthetic control is often a practical and rational path .
4. Diagnostics and robustness are mandatory for each strategy: pre-trend, density, first-stage power, bandwidth sensitivity, placebo tests.
5. Transparency (data source, preprocessing rules, code) and repetition are an integral part of the design.

As a result, when the hypothesis– variables –identification triad is harmonious, econometric evaluations become a reliable basis for scientific and practical decisions.

LIST OF REFERENCES USED:

1. Abadie, Alberto. 2005. "Semiparametric Difference-in-Differences Estimators." Review of Economic Studies 72(1): 1–19.

2. Abadie, Alberto, Alexis Diamond, and Jens Hainmueller. 2010. "Synthetic Control Methods for Comparative Case Studies." Journal of the American Statistical Association 105(490): 493–505.

3. Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. Mostly Harmless Econometrics: An Empiricist's Companion . Princeton, NJ: Princeton University Press.

4. Cameron, A. Colin, and Pravin K. Trivedi. 2005. Microeconometrics: Methods and Applications . Cambridge: Cambridge University Press.

5. Duflo, Esther, Rachel Glennerster, and Michael Kremer. 2007. "Using Randomization in Development Economics Research: A Toolkit." In Handbook of Development Economics , vol. 4, 3895–3962. Amsterdam: Elsevier.

6. Greene, William H. 2012. Econometric Analysis . 7th ed. Boston: Pearson.

7. Gujarati, Damodar N., and Dawn C. Porter. 2009. Basic Econometrics . 5th ed. New York: McGraw-Hill.

8. Imbens, Guido W., and Donald B. Rubin. 2015. Causal Inference for Statistics, Social, and Biomedical Sciences . Cambridge: Cambridge University Press.

9. Imbens, Guido W., and Thomas Lemieux. 2008. “Regression Discontinuity Designs: A Guide to Practice.” Journal of Econometrics 142(2): 615–635.

10. Pearl, Judea. 2009. Causality: Models, Reasoning, and Inference . 2nd ed. Cambridge: Cambridge University Press.

11. Popper, Karl. 2002. The Logic of Scientific Discovery . London: Routledge.

12. Rosenbaum, Paul R., and Donald B. Rubin. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects." Biometrika 70(1): 41–55.

13. Stock, James H., and Mark W. Watson. 2015. Introduction to Econometrics . 3rd ed. Boston: Pearson.

14. Wooldridge, Jeffrey M. 2010. Econometric Analysis of Cross Section and Panel Data . 2nd ed. Cambridge, MA: MIT Press.