https://doi.org/10.18222/eae.v34.9956_en

EXPERIMENTAL DESIGNS IN PUBLIC POLICY EVALUATION: USES AND ABUSES[1]

PAULO DE MARTINO JANNUZZI[I]
TRANSLATED BY: FERNANDO EFFORI DE MELLO[II]

[I] Instituto Brasileiro de Geografia e Estatística (IBGE), Rio de Janeiro-RJ, Brazil; *paulo.jannuzzi.br@gmail.com*
[II] Freelance translator, São Paulo-SP, Brazil; *feffori@gmail.com*

ABSTRACT

The present essay discusses the strengths and, above all, the limits of using experimental and quasi-experimental methods in evaluating public programs. It begins with a brief presentation of classical experimental design, its requirements and related concepts such as internal, external, and counterfactual validity. Next, it addresses the modalities of quasi-experimental evaluation design, which relax the requirements of the classical experiment. Two critical sections point out the ethical, political, and operational limitations of experimental and quasi-experimental designs in program evaluation and, subsequently, the political-institutional motivations for the resilience of this approach in spite of well-known and recurrent robustness problems.

KEYWORDS EVALUATION • PUBLIC POLICY • EXPERIMENTS • QUASI-EXPERIMENTS.

HOW TO CITE:

Januzzi, P. de M. (2023). Experimental designs in public policy evaluation: Uses and abuses. *Estudos em Avaliação Educacional*, *34*, Article e09956. https://doi.org/10.18222/eae.v34.9956_en

1 A preliminary version of this article was presented at the 46th Annual ANPAD Meeting, in 2022.

DELINEAMENTOS EXPERIMENTAIS NA AVALIAÇÃO DE POLÍTICAS PÚBLICAS: USOS E ABUSOS

RESUMO

O objetivo deste ensaio é discutir as potencialidades e, sobretudo, os limites do emprego de métodos experimentais e quase-experimentais na avaliação de programas públicos. Inicia-se com uma breve apresentação do desenho experimental clássico, seus requisitos e conceitos relacionados, como validades interna, externa e contrafatual. Em seguida, são abordadas as modalidades de desenhos quase-experimentais de avaliação, modelos que flexibilizam os requisitos do experimento clássico. Em duas seções de natureza crítica, apontam-se as limitações éticas, políticas e operacionais dos desenhos experimentais e quase-experimentais na avaliação de programas e, depois, as motivações político-institucionais da resiliência dessa abordagem diante de conhecidos e recorrentes problemas de robustez.

PALAVRAS-CHAVE AVALIAÇÃO • POLÍTICAS PÚBLICAS • EXPERIMENTOS • QUASE-EXPERIMENTOS.

DELINEAMIENTOS EXPERIMENTALES EN LA EVALUACIÓN DE POLÍTICAS PÚBLICAS: USOS Y ABUSOS

RESUMEN

El objetivo de este ensayo es discutir el potencial y, sobre todo, los límites del uso de métodos experimentales y casi experimentales en la evaluación de los programas públicos. Comienza con una breve presentación del diseño experimental clásico, sus requisitos y conceptos relacionados, como validez interna, externa y contra fáctica. A continuación, son abordadas las modalidades de los diseños de evaluación casi experimentales, modelos que flexibilizan los requerimientos del experimento clásico. En dos secciones de naturaleza crítica se señalan las limitaciones éticas, políticas y operativas de los diseños experimentales y casi experimentales en la evaluación de los programas, y después, las motivaciones político-institucionales de la resiliencia de este enfoque frente a conocidos y recurrentes problemas de robustez.

PALABRAS CLAVE EVALUACIÓN • POLÍTICAS PÚBLICAS • EXPERIMENTOS • CASI-EXPERIMENTOS.

Received on: DECEMBER 12, 2022
Approved for publication on: APRIL 11, 2023

INTRODUCTION

Experimental and quasi-experimental designs have a prominent role in the evaluation of public projects, programs and policies, as well demonstrated recently in the analyses of the medical safety and preventive efficacy of the vaccines developed – and other medications used – to combat the effects of the COVID-19 pandemic. Decisions in public health programs involve high risks, which can have rapid consequences far beyond predicted in concrete situations of social reality. Fortunately, epidemiological investigation protocols were developed a long time ago and enjoy strong consensus in the country's political and scientific communities, thus ensuring – until a few years ago – consistent and responsible decisions on public health policies in the country.

In other fields of public policy, such as the evaluation of educational and social programs, or even public health programs involving more complex interventions – beyond the development of vaccines, medications or clinical and therapeutic procedures –, the pertinence and applicability of these evaluation designs are still the object of discussion. For a segment of the community of evaluation practices, called randomists by Ravallion (2009), experimental research models – and, with some concessions, some quasi-experimental designs – constitute the gold-standard method in program evaluation, the only one that could attest the effectiveness and impact of a public intervention (Gertler et al., 2015). It is argued that, granted its application presuppositions, this research model can ensure more consistently the inference of causality between the intervention, its activities and products, and its effects. Following what is suggested in an "official" public policy evaluation manual issued by the Casa Civil da Presidência da República [Office of the Chief of Staff] (2018), evidence produced through this type of evaluation would be more robust for formulating and deciding on public policy, with a complementary – and not as valid – role being assigned to the results of evaluation through other approaches. The context of the primacy of fiscal austerity in recent years and the public repercussion of testing procedures for COVID-19 vaccines has certainly fostered this tendency, legitimizing the conviction about the experimental model as the "most scientific method" of evaluation of any governmental initiative in health, education or social protection. The fact is that the belief in the method remains vivid and efficient in capturing hearts and minds in universities, the public sector, and even the media in the country.

Recovering and systematizing arguments about the epistemological, ethical and operational limits of this evaluation approach, present in several contributions by national and international authors, is the goal of this essay. Starting by recognizing the potential of experimental and quasi-experimental methods in evaluation, the text discusses the "place" and "time", or contexts and moments, of their most

suitable and profitable use in assessing public policies and programs. The intention is to warn about the use, abuse and misuse of experimental methods for evaluating programs in the country, so as to avoid replicating the American experiences of the 1960s and 1970s, in which fragile – though supposedly "authorized" – evidence served to discontinue or delegitimize public programs newly implemented or which had not had sufficient time and resources to deliver what they proposed (Patton, 1990, 1997; Rossi, 1987). If public policies have political relevance and legitimacy by the purposes they are destined to – that is, they are socio-political constructions in a given context and society – it is necessary to evaluate them in a more consistent and responsible manner, according to the socio-technical models more appropriate for the process.

The essay begins with a brief presentation of classical experimental design and its requisites of sample randomization, laboratory/situational control, and clarification of causal relationship between variables. Related concepts, such as internal, external and counterfactual validity, are also introduced. Then, the modalities of quasi-experimental design are addressed. These models relax the requirements of the classical experiment, such as the randomization of treatment and control groups and the contextual conditions of "confinement". Thus, if on the one hand, in these modalities the causal inference loses the robustness of the classical design, on the other hand, evaluating the program becomes feasible by circumventing the indicated restrictions. The text concludes with two critical sections, pointing out the ethical, political and operational limitations of experimental and quasi-experimental designs in program evaluation and, subsequently, the political-institutional motivations for the resilience of this approach in spite of so many and well-known robustness problems.

A BRIEF EXPOSITION ABOUT THE USE OF THE CLASSIC EXPERIMENT IN PROGRAM EVALUATION

The classic experimental model in epidemiological research aims to investigate the structure and intensity of the causality between a consequent effect-variable (cure, improved health condition) and its determinant factor-variable (treatment, medication or vaccine). To that end, it is necessary to ensure the control of the experimental situation in laboratory and the use of two groups randomly formed from the same original population. One of these groups is submitted to the effects of the new treatment to be evaluated (the treatment group), while the other does not receive the new treatment, but can (and should) have the conventional treatment available (control group) (Imas & Rist, 2009).

The conduction of the experiment in laboratories or in a controlled context seeks to ensure the non-interference of other factors which might affect the study,

so as to make the effects – or non-effects – of the new treatment more evident. The random assignment of people to form the treatment and control groups seeks to ensure that, since these individuals are from the same base population, the two samples are probabilistically equivalent or similar in their characteristics. Therefore, three conditions are required for conducting an experiment: the supposed cause or explanatory factor (treatment, medication or vaccine) to be applied/tested during a certain period; control over the experimental/laboratory situation throughout the period of analysis; and random selection of the people to form the two groups – treatment and control. In a study where one of these conditions is violated, this cannot be called an experiment.

A hypothetical example inspired by the tests of the Salk vaccine against poliomyelitis (Cano, 2004) can help understanding the experiment logic. Let us suppose, for example, a clinical trial – in fact, a clinical experiment – for evaluating the efficacy of a medication – CoroX – against a new disease – CoroV – which is highly contagious and has therefore affected the health of thousands of people and caused the hospitalization of many of them in serious conditions. Thus, there is a specific factor or variable to be evaluated in terms of its impact: whether CoroX cures or mitigates the adverse effects of CoroV. With approval from the Comitê de Ética em Pesquisa [Research Ethics Committee] (CEP) and authorization from the Agência Nacional de Vigilância Sanitária [National Health Surveillance Agency] (Anvisa) for testing the medication, the preparations for the research can begin.

To evaluate the CoroX through an experiment, it is necessary to ensure the situational/laboratory control: to choose hospital with appropriate facilities for conducting the research and several patients hospitalized due to CoroV for effectively testing the medication. It is supposed that there is significant number of hospitalized patients with different levels of severity and lengths of stay in a given hospital. The last condition for the experimental research is the creation, through a random draw, of two groups of patients: the treatment group, which will receive regularly the medication CoroX; and the control group, which will be treated with CoroY, a product with a similar appearance, but without the active ingredient to be tested. It is a placebo, necessary to make each patient feel that they are in treatment, since extensive medical literature points out that psychological factors, such as being under medical care, especially in a cutting-edge study, can favor a patient's health. For medical and experimental record, tests are performed in each patient before treatment begins.

It is worth noting that, apart from CoroX, the two groups receive the same available treatment. By the principles of medical ethics, it is not acceptable to deprive a group – the control group – of the best conventional and already tested treatment available. Thus, it is not a matter of evaluating CoroX in relation to no treatment at

all, but rather to the conventional treatment available. The random assignment of patients in one of the groups should be, preferably, though a double-blind procedure: not only do patients ignore whether they are taking CoroX or the placebo CoroY, but also the nurses and doctors in daily contact with them ignore the groups to which they belong. This eliminates occasional personal conflicts and biases in treatment or in recording patients' health evolution. It also annuls occasional attempts to interfere – by meritorious reasons – in the experiment's design, like pressuring for the administration of CoroX to all patients, when it is perceived that the medication seems to improve the health of some of them. Certainly, these ethical and human conflicts are not easy to deal with, but are addressed by the pragmatic ethics of health experiments.

As days pass and the treatment advances along with the monitoring of each patient's health, the moment is progressively configured for the test or battery of tests for evaluating the condition of cure, improvement or non-improvement of patients. The moment of this test is "summative" and critical: if too early, it may not indicate the effects under processing; too late, patients in the experiment may by be in an even more severe state and die. The duration and intensity of the medication dose are critical variables to be analyzed and may require additional experiments.

At some appropriate point, it is necessary to evaluate whether or not CoroX has caused an impact by comparing the summary measures of average health status for both groups. As illustrated in Table 1, based on Rossi et al. (2004), the evaluation of CoroX's efficacy depends on the difference between T2 and C2, measured from the health status of both groups – treatment and control – at the evaluation moment: if T2 is significantly greater than C2, then CoroX appears to make a difference in patients' improvement. In the most general case, the experiment's evaluation should be based on the significant difference between T2-T1 and C2-C1, that is, whether the average health status of the treatment group showed a greater improvement than that of the control group (if T2-T1 > C2-C1). It should be "significant" as it refers to an experiment with samples of the patient population, and not all of it. Adopting a standard level of statistical significance, if the eventual difference of the improvement measures between the groups is statistically different from zero [(T2-T1)-(C2-C1) > 0], then there is strong evidence that the treatment with CoroX has an impact on the cure or improvement in the health of CoroV patients. If, on the contrary, no significant difference is observed between the two groups for health improvement, then there is no evidence that the medication CoroX is effective.

TABLE 1
Evaluation according to the classical experimental design

T1 ----------→ T2
C1 ----------→ C2
Measure before treatment begins — Measure after enough treatment time
Where T: Treatment group, C: Control group
Groups defined by random assignment and under "laboratory" control.
If T2 > C2 or, in the most general case, if T2 – T1 > C2 – C1,
then there are indications that the treatment is effective.

Source: The author, based on Rossi et al. (2004).

The intuitive idea underlying this experiment is that, since both groups were exposed to the same "laboratory" conditions (inpatient care in the same hospital) and are very similar (as they are samples of the same initial set of CoroV patients), the difference between them would be a consequence of the fact that one of them had access to the treatment, while the other did not. The difference in health status evolution between two similar groups of individuals kept under the same external context should be ascribed, since there is a difference in the treatment the groups are receiving. This characteristic or property of the experiment is denominated internal validity.

More precisely, the internal validity of a research that seeks to attribute a causal relationship between two variables is the certainty ensured by its methodological design that the independent variable – or the factor under test – is responsible for the observed impact (positive or not) on the dependent variable – or the observed outcome (Imas & Rist, 2009). As clearly pointed out by Cano (2004, p. 29, own translation), "internal validity is the degree of certainty that the effect on the dependent variable of the experiment was caused by the independent variable of the experiment". It is a matter of attesting to what extent it was the researched cause, and not another factor, which produced the observed effects.

A design enjoys high internal validity when there is great certainty that the key variable under evaluation – the medication CoroX, in the example – is the only vector responsible for the observed outcome: improvement or non-improvement in the CoroV patients who were treated with the medication. By their methodological design – randomization of groups and their submission to identical external conditions –, experiments enjoy high internal validity.

Internal validity is not a dichotomous property (it is there or not), but one that is evaluated in graded terms (high, medium and low). Thus, other not strictly experimental research designs, such as the comparison of CoroV patient groups from different hospitals, or the comparison of the same group of patients before and after the medication was administered, have less internal validity. That does not mean that such studies – typical modalities denominated quasi-experimental – do not enjoy any internal validity, but rather that the certainty that the use of CoroX

is responsible for an eventual improvement observed is smaller than in the case of the reported experiment. After all, the teams at both hospitals can be different, as well as the equipment and facilities, or the improvement over time can derive from a positive patient reaction to the set of therapeutic procedures, and not to CoroX alone.

If internal validity concerns the certainty about the intrinsic causal relationship – or lack thereof – between two variables, the external validity of a research refers to the certainty that its results can be generalized to other contexts, populations or methodological variations in the measurement of variables. Relying again on Cano (2004, p. 29, own translation), "external validity indicates the extent to which the causal inference proposed by the experiment can be generalized to other moments, places, populations and forms of measuring the variables in question – both independent and dependent ones".

As commented earlier, the study evaluating the efficacy of CoroX has high internal validity, ensured by the experimental design adopted. However, one cannot ensure beforehand the external validity of its results, i.e., that CoroX will be effective to treat any CoroV patient, in any hospital, and even less that such medication serves to prevent the contagion of the disease. After all, the experimental situation was quite peculiar: a group of patients under inpatient care, in a particular hospital, in a certain city. Would the treatment with CoroX be effective for patients in hospitals with a smaller medical team, less infrastructure, subject to equipment problems and working beyond capacity? In locations which are colder or warmer, wetter or drier, more or less populous?

In order to expand external validity and ensure the generalization of clinical experiments, multi-center studies are conducted, i.e., similar evaluations for diverse population groups, regions, contexts of treatment and severity of the disease in question. Thus, these studies seek to reproduce the innumerous concrete and real situations encountered in public health, and no longer in the controlled context of technical-scientific research. Ideally, to ensure the study's internal and external validity, the sample of people randomly selected for the treatment and control groups should be an approximate portrait of hospitalized patients or even of healthy individuals

As suggested by Gertler et al. (2015), in order to allow generalizing the impacts identified in experiments, the sample should be representative of the eligible population, though with treatment and control groups formed through randomization processes.

> An impact evaluation can generate internally valid impact estimates through the random allocation of treatment; however, if the evaluation is carried out in a non-random sample of the population, the estimated impacts may not be generalizable to the population of eligible units. Conversely, if the evaluation

> uses a random sample of the population of eligible units, but the treatment is not randomly selected, then the sample will be representative, but the comparison group may not be valid. (Gertler et al., 2015, p. 55, own translation).

But it is not always so simple to ensure simultaneously the internal validity (causal inference) and external validity (populational inference) in multi-center clinical experiments or experimental evaluations of public programs. Ensuring one might hamper the other (Cano, 2004). Greater experimental control to afford greater causal inference power can mean greater artificiality of the analyzed context, thus limiting the generalization of results to more realistic contexts; broader population-representative samples can fragilize the experimental presuppositions of contextual control (and the study's internal validity).

In the experimental evaluation of public programs, balancing both types of validity acquires even greater criticality. How to ensure that a program that is well evaluated in artificially built circumstances can repeat the success in normal situations across the country, with all the heterogeneity of public services and facilities? If internal validity is an important attribute to be ensured in evaluation, in order to attest or not the program's impact, "what good is a full guarantee about a causal inference which cannot be applied beyond the concrete context it was generated in?" (Cano, 2004, p. 31, own translation). Therefore, it is necessary to seek balance between internal and external validity in experimental evaluation, or in any research design.

THE MAIN MODALITIES OF QUASI-EXPERIMENTAL EVALUATION DESIGNS

Quasi-experimental designs are actually much more common than experimental ones in the evaluation of programs whose target population are people, families, companies or institutions. Though with lower internal validity, they are more feasible in the Brazilian context. In the concrete reality of public administration, public program impact evaluations with strictly experimental designs are less feasible than normally recognized in some evaluation manuals. There are situations where the term "natural experiment" is improperly used for the evaluation of a program when, in fact, it is not a classical experiment like the one presented here, but rather a quasi-experimental modality. This is the case when, due to implementation schedule issues, some families or individuals are served before others who are equally eligible for the program. Supposing that the access of families or individuals occurs in a "natural" way, without personal, family or regional priority as criteria, the process would emulate a pseudo-randomization – rather than an authentic randomization – in the formation of the treatment and control groups, which precludes classifying that evaluation as strictly experimental.

As in experiments, quasi-experimental evaluation is aimed at determining whether the observed effects – impacts – on a group of beneficiaries or users of a program derive from activities, products and components developed in it, as well as estimating the dimensions of that impact. What differentiates quasi-experiments from classical models is generally the non-observance of the random assignment of treatment and control groups, or the lack of contextual control (and of a control group), and therefore the relaxation regarding the interference of other external factors in the production of a program's effects (Cano, 2004; Rossi et al., 2004). Either situation ends up violating the presuppositions of the classical experimental model, affecting the assertiveness in the attribution of the program to differential impacts observed in its beneficiaries in relation to the comparison group (and not a control group, the term used in the experimental case, as seen earlier).

In quasi-experiments the treatment and comparison groups end up being defined by non-random or, at best, pseudo-random processes, as in the previously described situation of the "natural experiment". In this case, due to delays in or the implementation schedule of a public program, or to a deficient coverage of the program's target population, it is possible to artificially create comparison groups from the analysis of the groups already included. Families, people or companies eligible for the program, even if not yet included in it, can serve as sample units for the comparison group, from which measures can be collected for purposes of comparison with the treatment group. Additional care, with greater or smaller methodological refinement, can ensure that groups become more comparable, such as the application of PSM (propensity score matching) calibration factors, correcting the differential effect of socioeconomic and demographic characteristics between both groups, before and/or after the "treatment" period. Anyway, the non-randomization in forming the treatment and comparison groups introduces what the literature of the area classifies as a selection bias in quasi-experiments. Non-random samples of the same population have some bias, whether more flagrant or less transparent. Non-randomized treatment and comparison groups have a selection bias. Even if their elements – families, people and companies – are part of the same target population eligible for the program, the pseudo-randomization strategy and the palliative calibration resources cannot ensure that both groups are equivalent in every possible dimension that might affect the outcome at the time of participation (or non-participation) in the program.

Among people, families from the same socioeconomic stratum or sociocultural context, or companies and institutions with the same sector or size characteristics, some have more resources, information, initiative, interest in or urgency to enroll in a public program. These motivations are hard to be objectively quantified, escaping the possibilities of calibrating and equalizing the characteristics of the groups enrolled and already participating in the program and of those not participating

but desiring to do so. The fact is that this motivation can interfere with the outcomes and "inflate" the estimated impact of participating in a program, or, in the view of Gertler et al. (2015), lead to estimating a false impact.

However, outside manuals and the academic modeling of the world, reality is never so simply dichotomous (right/wrong, true/false, etc.). Thus, in the complex reality of evaluation research, in the same way as internal and external validities are measured on an ordinal scale (high, medium, low), impact estimates can be more or less consistent or robust, the latter being the term most used. However, the robustness of an impact estimate does not depend only on the efforts to avoid participant selection bias at the beginning of the evaluation design, but also on ensuring a good conduction of each evaluation step, which is not trivial.

Naturally, all possible actions should be taken *ex ante* to minimize the selection bias, respecting the ethical limits and political risks mentioned earlier. Likewise, it is necessary to recognize *ex post* biases deriving from the concrete operationalization of the evaluation design and its effects on the final composition of both groups – treatment and control. Nor can it be forgotten that experiments or quasi-experiments are always subject to some bias of selection of the target population that is the base for the evaluation study, i.e., its representativeness in relation to the population potentially eligible for the program. In more precise terms, there is the inevitable selection bias derived from non-randomization – which affects the evaluation's internal validity – and the eligible population selection bias – which affects the design's external validity. Estimates of the program's effective impact in its day-to-day operationalization will be more or less consistent depending on the combination of strategies of simultaneous mitigation of both types of biases.

As observed earlier, while they can also suffer from various of the earlier-mentioned operational problems and have a lower validity than experiments, quasi-experiments have greater feasibility, since the ethical and political-institutional restrictions are better addressed. They seem to provide a methodological alternative of impact evaluation with a balance between internal validity, technical rigor and practical feasibility.

In the conceptualization proposed by Imas and Rist (2009), there are modalities of quasi-experiments, five of which are of special interest in this unit, classified into two categories: evaluation designs with multiple measures for the same group, without comparison with another group; and evaluation designs using similar comparison groups (Table 2).

Generally, quasi-experiments without comparison groups are models with a lower internal validity than those relying on comparative groups for counterfactual estimation, i.e., for estimating the situation the group of beneficiaries/users of the program would experience over time if they were not covered by it.

TABLE 2
Experiment and modalities of quasi-experiments in impact evaluation

Classical experimental model	T1 C1 T2 x C2 Estimated impact = (T2-T1) – (C2-C1)
Quasi-experiment without a comparison group	
Design with before-and-after evaluation	T1 x T2 Estimated impact = T2 – T1
Discontinuous historic series	T1 T2 T3 x T4 T5
Longitudinal design	x T1 T2 T3 T4 T5
Quasi-experiment with comparison groups	
Post-treatment evaluation with a comparison group	T2 x NC2 Estimated impact = T2 – NC2
Design with pre and post-treatment evaluation	T1 NC1 T2 x NC2 Estimated impact = (T2-T1) – (NC2-NC1)

Source: The author, based on Imas and Rist (2009).

Note: T: treatment group, program beneficiaries; C: control group (randomized); NC: comparison group (not randomized); X: treatment, program.

The evaluation with groups with pre and post-treatment measures is the model whose design is closest to the classical experimental model. Supposing the similarity or pseudo-randomization of treatment and control groups, this is an impact evaluation model more sophisticated and with greater internal validity than the others. The impact can be estimated as the difference of the differences of the pre and post-measures for each group. The resulting balance is the impact [(T2-T1) – (NC2-NC1)]. Such process would ensure more consistent impact estimates, since biases of the same group would supposedly be eliminated by the differences of the measures of results before and after.

While this is a model advocated as desirable – and, for some communities, as the only impact evaluation model acceptable when an experiment is not possible –, it is less frequently applied on a large scale to analyze federal programs in Brazil, due to its cost, operational complexity, and time length between planning, two-wave data collection and the production of results. Nevertheless, one of the examples to be highlighted in this respect is the impact evaluation of the Bolsa Família program (Jannuzzi & Pinto, 2013). The quasi-experimental methodological design was used to capture impacts specific to the program on various socioeconomic dimensions, based on the collection of data on the socioeconomic situation of beneficiary and non-beneficiary population groups at two moments (2005 and 2009). In order to ensure internal and external validity for the evaluation and preserve it from possible questionings about different programmatic interventions in municipalities of its sample, the addresses – and municipalities – of the households interviewed on the first round of the survey was not informed by the Center for Regional Development and Planning (Cedeplar/UFMG), responsible for the field study, to the Secretariat for Information Evaluation and Management of the Ministry of Social Development and Fight against Hunger.

This decision to preserve statistical secrecy, or rather, to disidentify the sample available to the ministry, was kept in the second edition, conducted in 2009 by the consortium formed by the International Food Policy Research Institute (IFPRI) and the Datamétrica Institute.

As with the 2005 data collection, the evaluation investigated a broad range of dimensions of the families' living conditions, including housing conditions, demographic and education characterization, participation in the job market, income, perception about social programs and events focusing on the health and anthropometry of children younger than 5 years old. The research design allowed obtaining additional evidence about the program's impact, which adds to the wealth of evidence produced by several other studies. However, it would be an error to say that, because of the more sophisticated design of the study, these results were to have more public repercussion or technical recognition by the management community. This is not what happened, as even other studies conducted in the program had low appropriation by the media and society.

A CRITICAL PERSPECTIVE ON THE MYSTIFICATION OF EXPERIMENTAL EVALUATION DESIGNS

Contrary to what is defended by Gertler et al. (2015) and the epistemic communities of randomists – thus critically named by Ravallion (2009) –, experimental designs are poorly applicable to concrete cases of public policies for ethical, political and operational reasons. They would even be a waste of public resources in many situations where their application is insisted on, as Moral-Arce (2014, p. 40, own translation) correctly puts it:

> Unfortunately, stakeholders, at various levels, believe that routine impact evaluations can (and should) be conducted for all programs. Paradoxically, this insistence on trying to make impact evaluations in a systematic manner can lead to the undesired result of resources (which are limited) being wasted on trying to make an evaluation of this kind.

Firstly, to apply the classical experimental design in public policies, there are several non-trivial ethical problems of how to choose and justify who will be a beneficiary and who will be out of the program, a question that that in medical research practice has been addressed by ethics committees. Such bodies assume the responsibility for, and the risk and legitimacy of these choices for the sake of scientific development and the promotion of cure and health, but they ensure to patients of the control group the best treatment available at the time. In the field of public policies, the situation at hand may be one of evaluating the effect of a new program compared to no public service available.

In the Brazilian case, the selection of beneficiaries of programs through a draw is still highly arguable from the ethical standpoint. In a society marked by high inequality between regions, colors/races, socioeconomic conditions, and by social iniquities such as poverty, hunger, child labor, it does not appear that "the ethical promise of scientific development" can silence or prevail over the "ethical commitment to human dignity". Combating inequality, ensuring basic social rights and promoting human dignity are public values and principles present in constitutional and sub-constitutional legislation in the country. Not providing access to a public transfer payment program for a poor and eligible family, to a cistern for water storage for consumption and production for a family in the country's semi-arid region, to an opportunity to enroll in a professional training course for an unemployed worker, to social protection services for a mother or family with violated rights, when there is the possibility to serve them, by no means appears compatible with any ethical-social or ethical-public reasonability. This is not a matter, as it is in clinical experiments, of providing access to a "new treatment" to some draw-selected (or lucky?) few, ensuring to the others a "conventional treatment". Given the intermediate and incomplete status of public policies in the country, adopting a classical experiment in the implementation of a new program means, once again, to reproduce the social inequality and iniquity of public action, providing to some the package of benefits and services, and to others, in a probably similar condition, no access to these public goods or services.

There certainly are arguments that relativize this stance. A "pro-social dignity ethics" is counterposed with a "pro-spending efficiency ethics" which, in the interpretation of the authors of the chapter "Impact evaluation" of the manual *Avaliação de políticas públicas* (Public policy evaluation) released by Casa Civil in 2018, would, in the medium term, lead to the former:

> The questioning of whether it is ethical to use the random selection method should also be placed in a broad context about the efficient use of public resources. Many policies do not have their impacts evaluated and continue to operate on the assumption that they reach the desired effects. However, it is necessary to recognize that we can only trust the effects of a treatment when it has been submitted to tests with high scientific rigor. Oftentimes policies considered effective do not show any impact, or present even impacts contrary to what is desired after being evaluated with robust methods.
>
> The lack of rigorous evaluation therefore hinders determining whether resources are being wasted that could be more efficiently employed, even to the same goals. In sum, this discussion shows that the critique that randomized impact evaluations are unethical should be placed in a broader context – one that is therefore capable of considering the flaws resulting from the selection

> of beneficiaries in alternative scenarios, the reality of implementation of public policies, their territorial budget limitations, as well as the efficient use of public resources evaluated in the most robust way possible. (Casa Civil..., 2018, p. 271, own translation).

But in the same chapter, a little before the quote above, the authors recognize that there are situations where random selection would be poorly justifiable and politically risky.

> To ensure that randomized evaluations are implemented in an ethical manner, it is necessary that at least two aspects are met, such as priority for those with special needs and draw-selection transparency. Should there be groups identified as having special needs who require priority by the policy, these should not be excluded for the sake of evaluating the impact of the intervention. In the context where the method of random selection is employed, the members of these groups should be benefited without taking part in the randomization. As for transparency, it is best practice to use public draws after enrollment, attended by those who enrolled. This kind of open and clear procedure ensures that the selection is made not only by offering equal opportunities of entry to all, but also preventing criticism against the policy's managers for favoring particular persons or groups. (Casa Civil..., 2018, p. 270, own translation).

In Brazil, as already suggested by Cano (2004), the provision of access to public programs and services through "draws" seems to remain little accepted, whether by the population or the body of public servants. The author comments that, in countries where the principle of experimental program evaluation is accepted, this "took decades of persuasion" (p. 24, own translation).

Ethical questions aside, always in dispute, in operational terms, there is from the outset a problem of scale in experimental evaluation: in public policies, the dimension addressed is not that of hundreds or a few thousand people as in clinical trials or social projects; they involve tens and hundreds of thousands, if not millions of people to be potentially served. Representative samples would be inevitably large and costly in contextual control and treatment. Unless it is feasible, as noted a little earlier, to ensure external validity with a sample size that is justifiable in terms of costs, the experimental model should not be employed

But the conflict between internal and external validity is just one of the several concrete problems facing the application of the experimental model for public program evaluation. The literature about experimental designs is rich in instances of concrete difficulties and problems with the method which affect its internal and external validity, such as those indicated by Cano (2004) and Imas

and Rist (2009): previous trend effect; effect maturation; loss of impact measure discrimination; regression towards means; sample mortality; interaction with other factors; interference with other policies; etc. The difficulties to communicate results and the high implementation costs are other problems pointed out. Ravallion (2009) adds another one: evaluative experiments test a specific hypothesis and do not show which component of a public program is hindering its full implementation.

It is also necessary to point out an epistemological dilemma in the experimental design: how to ensure that complex public intervention logics, involving several actors, in so very different contexts, can be modelled in variables with regular and accurate measurement in dimensioning the efforts and effects of policies and programs? How to ensure that program X, its component Y or Z, in a long and complex intervention chain X-Y-W-Z is the causal factor of Z? Moreover, in a context of simultaneous implementations of various public programs, there is a high probability that other programs or factors are contributing to produce the impact. That is what Bamberguer et al. (2016) argue in a document about the evaluation guidelines of the Sustainable Development Goals Agenda for UN Women:

> In small projects with a low level of programme complexity, relatively simple institutional arrangements and a low level of contextual dependence, it is possible to trace and evaluate a direct causal relationship between a programme intervention (e.g., drinking water, scholarships for girls to attend secondary school) and the intended outcome (e.g., lower rates of diarrhea, higher rates of girl's enrolment). As programmes become more complex in terms of the three previous dimensions, the number of inputs increases (often operating differently in different communities or regions), the number of intended and unintended outcomes also increases, and the influence of different stakeholders and institutional arrangements becomes more complicated, as well as the number of contextual factors. Consequently, it becomes increasingly difficult, or in many cases impossible, to determine direct causal relationships. (Bamberguer et al., 2016, pp. 46-47).

The fact is that the attribution of impact to a program or a component thereof is increasingly difficult in contexts where public policies are designed to achieve multiple goals (even with different emphasis between them). Thus, rather than measuring marginal contributions of programs, would it not make more sense to measure joint or combined effects?

There are yet operational difficulties to conducting an experimental evaluation: how to prevent beneficiary drop out, the entry into the treatment group of other individuals who were part of the control group, and still ensure that the measured effects are not affected? Experimental evaluations of programs involve

field data collection for at least two rounds or waves, so as to obtain pre and post-program measures for the treatment and control groups. Depending on the time length between both survey waves, one group or the other can lose individuals or families due to address change, migration to another location, death or refusal to participate (especially among non-beneficiaries). If the loss of sample units – or attrition – is differentiated by both groups, the internal validity may be weakened, since hypotheses might be raised about other variables interfering with the process. Even external validity could be weakened if the final sample no longer aligns with the program's reference target population. During the interval between both waves, beneficiary families or people may leave the program or be excluded from it. Or, as is usually more common, families or people from the control group can become beneficiaries of the program, thus reducing the group's sample size and, again, decreasing the experiment's internal validity. This fact interferes with the time of exposure to the program, a key variable for impact intensity. Families or individuals who are beneficiaries of a program can have their access to another program facilitated by policymaking decisions or by gaining increased knowledge about public policies. Another setback for internal validity. In real life, contextual control in program evaluation is much less effective than it is possible in laboratory tests of medications and vaccines. Finally, there is the possibility that the sample in the second survey wave presents treatment and control groups with significant biases, eliminating all guarantees of internal and external validity that the random selection afforded at the beginning of the evaluation.

There are also methodological challenges in choosing the best measure for capturing the impacted dimension. Does what is assumed to be an impacted dimension keep, by the program's logical design, a close connection with its actions, products and services? Is the effect to be measured a concrete outcome of the program, ensured only by the program, or a desirable or potential effect requiring other presuppositions or actions not foreseen in the program design? Should the impact be measured on the beneficiaries, their families, their community, or the wider society?

Even if all these problems could be circumvented, there would still remain one of a practical nature: if the potential effects of the program, as measured in a particular variable, are not high, then the samples of beneficiaries served and in the control group would have to be considerably large so that the statistical tests might be accepted without hesitation (Rossi et al., 2004). The anthropometric indicators considered in the impact evaluation of Bolsa Família are a clear illustration in this respect: since the program's monthly financial transfer is not enough to bring about significant changes in family diet, the differences of height and weight between children from families in the program and those not covered by it are of little

significance. If the sample were larger, probably the program's impact would have been verified not only in the body mass index, but also in children's average height, a more refined analysis indicator (Jaime et al., 2014).

Regarding the use of significance tests in social research, Sawyer and Peter (1983) comment that it is illusory to consider them as strictly objective procedures when the researcher is given the opportunity to alter, to their liking, the several parameters that can determine the significance or non-significance of an association between variables. Increasing or decreasing the size of the sample the test is applied to, choosing a one-tailed or two-tailed test, changing post hoc the level of significance, are subjective procedures – and not necessarily reprehensible ones – which are behind the apparent formality and mathematical accuracy of the significance tests. There is even a significant movement of researchers who propose to abolish significance tests, for considering that they suggest a "robustness" in the findings which they could not actually provide (McShane et al., 2019).

In the applied literature there are many *ad hoc* solutions for several of these concrete problems, from the calibration of samples of the two groups to instrumental variables to other econometric solutions which are not so consensual. But it is necessary to be attentive to the existing limits in the adoption of procedures to correct non-responses, sample unbalancing, sample losses in field surveys. The alleged differential rigor of the experimental model in relation to other program evaluation designs may stay in the evaluation plan, being gradually lost at each concrete step in the execution. Perhaps, in some cases, the most suitable methodological solution – and the most honest one from a scientific standpoint – is to recognize that the evaluation has ceased to be experimental to become a quasi-experiment.

WITH SO MANY ROBUSTNESS PROBLEMS, WHY DO THESE EVALUATION DESIGNS RESIST TO TIME AND CONTEXTS?

The belief that experimental impact evaluations and their variations constitute the gold standard is reinforced, in a "self-referenced" circle, by multilateral development banks and other communities of social project funders. These institutions, usually constituted by teams with a markedly disciplinary and positivist academic background, with little knowledge of program management design and practice, reinforce this perverse logic professed by this epistemic community: they only allocate resources to initiatives where the managers undertake to follow the previous impact evaluation rule book, whatever the nature of the intervention, operational feasibility of the design, or ethical principles to be obeyed.

This is what La Rovere (2014) discusses in the context of environmental policy evaluation, where the investigation of marginal contributions of initiatives and the

separation of investigated units into treatment and control samples are operationally unfeasible. And the author exposes the functioning of the funding-method-funding cycle in projects and programs:

> . . . yet pressure arising from multiple sources (donors and evaluation fora) towards the perceived higher rigour achievable through quantitative approaches and attribution is being reapplied on impact assessment and evaluation practitioners. This demand is stimulated (or often enforced) by major donors insisting that a quantitative approach is the only credible one. These influential donors are almost always located in the same places (i.e. countries, cities and often intellectual circles) as the academic institutions where such tools are being promoted. (La Rovere, 2014, p. 285).

Along the same lines, an extensive study dedicated to understanding why experimental models had a second ascent wave from the 2000s onwards, after the decline in the 1970s, explains that:

> Our argument is that the contemporary expansion of RCTs can only be understood in the context of two independent transformations that then became linked to one another in elective affinity. On the one hand, the field of foreign aid has been profoundly transformed by the entry of a new set of actors: large, private foundations with a global ambition and a new managerial style (the so-called "philantrocapitalism". . . . On the other hand, the field of development economics has been transformed by the rise of behavioral economics, with their emphasis on cognitively plausible models of human actors and the use of "nudges" to channel actors in an economically rational direction. . . . The success of randomistas must be understood as a function of their capacity to exploit the elective affinity between these two transformations, with RCTs serving as the "hinge" between the "inked ecologies" of the economics profession and the field of development aid. . . . The success of randomistas, we shall see, depended on their capacity to effectively change what is a "field experimente" and what it means to evaluate development policy, a change for which the privatization of foreign aid provided a hospitable ecology, while behavioral economics provided a conceptual model and a set of tools. (Leão & Eyal, 2016, p. 3).

The authors therefore attribute less to the propagated robustness of the experimental method and much more to the conjunction of interests – elective affinities[2] – of a new school of economic thought in a new context of development

2 Simply put, "elective affinity" is a sociological concept recurrently employed to illustrate situations of convergence of interests between actors or a structural correspondence between them.

project funding. The philanthrope-funders of social projects, whether by their education or need for "objective" procedures, were of the conviction or had been convinced that experimental evaluation would meet these goals well. The community of economists of this thew behavioral school saw the opportunity to restore the toolkit produced a few decades ago and repack it as the newest and most robust evaluation methodology, the gold standard to be used, to the detriment of the other approaches. As it happens, though, according to these authors, this evaluation approach had ceased to be used precisely for its inadequacy for the implementation problems and complexity of public programs at the time.

The mystification of this design in program evaluation is owing, to some extent, to the origin of evaluation studies in the investigation of public health and education programs, where such models can more easily become feasible – by the "laboratory" simulation conditions in classrooms or through the tradition of clinical trials for the treatment of diseases (Leeuw, 2010). Another explanatory factor is the circumstantial hegemony of the quantitative models from natural sciences in the debate about the scientificalness of research methods in American social research in the 1960s – a moment of expansion of evaluation studies in that country (Jannuzzi, 2018). Indeed, the classical book *Experimental and quasi-experimental designs for research*, published in 1966, strongly influenced evaluation training and practices in American universities, in part due to the elegance and internal validity for causality analyses. In this period, despite warnings about the difficulties to replicate laboratory conditions in the context of social program operation, "the elegance and precision of the experimental method led most program evaluators to see it as ideal" (Worthern et al., 2004, p. 116, own translation).

The criticisms that followed in subsequent decades about ethical aspects, operational feasibility and generalization power of the results of experimental designs – and their quasi-experimental approximations, whether in academic research or in program evaluation studies –, the incorporation of evaluators from several social science disciplines – anthropologists, sociologists, communicologists, etc. – and the more rigorous formalization of qualitative investigation approaches, more appropriate for the complex and little structured problems of social reality, ended up consolidating the perception, in the community of evaluators in the United States, that evaluation studies require some methodological eclecticism, integrating quantitative and qualitative methods.

Experimental and quasi-experimental public policy evaluation designs involve not only ethical questions that are difficult to circumvent – such as choosing who takes part in the control and treatment groups – but also operationalization problems that are far from trivial – such as the "isolation" of both groups over time and ensuring the "isonomy" of the other contextual conditions. In a reference to outdated views about

the production of scientific knowledge, they declare themselves "politically neutral" and "scientifically attested". They forget that the attribution (or delegitimization) of identified effects on a population to the components of a program often depends on many choices regarding statistical tests, significance levels, sample size and characteristics, the presuppositions in relation to data distribution properties. In many of these studies there is no discussion about the statistical power of the tests used or about the analysis of residues after estimating model parameters. Even less common are more exhaustive analyses about the potential biases introduced in the estimation of direction and intensity of the impact (or non-impact) by the calibrations of the treatment and control groups through propensity score matching. The impression one gets from many works is that causal relation hypotheses are accepted or rejected more by convictions than actually by "full evidence".

Moreover, the choice of the analysis sample, which allows dodging ethical or operational imperatives in separating treatment and control groups, is at times highly particular, suffering from the critique they impute to the samples of other studies they repute as having "less scientific" designs. In other words, these are also samples with a representation (and why not selection?) bias, where external validity may be compromised, in order to ensure internal validity conditions (randomness or quasi-randomness in defining both groups). If the selected samples boost the internal validity of the evaluation research (and the relationship of attribution between cause and effect), it must be acknowledged that this is often done to the detriment of the representativeness of programs' target populations and the hard and concrete reality of program implementation in complex environments.

Experimental designs certainly have relevance and application in public policy evaluation. Its employment as an impact evaluation strategy requires a series of reflections about the timing, costs and expectation of results. Classifying them as gold standard is not only a mistake in terms of general prescription in relation to the different evaluation questions possible for a particular public policy, but it is also little responsible for the ethical and political implications required for its realization.

There is no single way of doing science or a single "gold-standard" method for producing and legitimizing knowledge. Nor do "absolute and undisputed truths" exist in sciences, let alone in social sciences and in evaluation (Gussi & Oliveira, 2017). Such "truths" are what some religions look for; in social sciences and evaluation, the search is for consistent findings and interpretations about what is analyzed, or from a "Latourian" perspective, doing science is building persuasive and convincing narratives about the analyzed findings, narratives that are recognized as legitimate and reasonable by the epistemic communities one belongs to.

A similar critical position is held by Chianca (2015, p. 28, own translation), expounded in an interesting study, whose final considerations are worth quoting in full:

> a. To say that a single method is the gold standard is like saying that a single medication is the best there is. You need to ask yourself always whether it is the best one for what kind of health condition.
>
> b. The true gold standard is the one that can establish a causal argumentation that is consistent and within the need for precision required by the evaluation context, based on correct and robust evidence which, at the same time, supports and critically tests that argumentation.
>
> c. Choosing a single method is not desirable. By far, the best evaluation or research designs use the principle of critical multiplism . . . which means employing a combination of methods, the strengths of one offsetting the weaknesses of the others and vice-versa. Every method has its limitations. Therefore, relying on just one of them is an unsuitable practice.
>
> d. There is one final critical point to be considered. One should not choose only the method or set of methods that seems technically more appropriate for a situation. One works with real-life situations, with resource limitations. Therefore, it is necessary to choose causal inference methods which are more cost-effective in order to respond in a sufficiently robust way to our evaluation questions.

As defended by Weiss (1998), Vaitsman and Paes-Sousa (2009), Batista and Domingos (2017) and several other authors quoted here, such as Imas and Rist (2009), Moral-Arce (2014) and Bamberguer et al. (2016), policy and program evaluations have much to gain in quality and consistency by using complementary approaches of quantitative, qualitative, experimental and quasi-experimental methods.

FINAL CONSIDERATIONS

Experiments are little feasible in Brazilian reality, even with the pressure of two decades of multilateral development organizations to make them more regular in the evaluation of programs. Ethical and political questionings about a possible access to public programs by means of draws still seem insurmountable to society and public managers in the country. In the scenario of structural inequalities and iniquities in Brazil, it does not seem sensible to surrender to methodological inclinations when there are substantive eligibility parameters and republican criteria for social prioritization to guarantee not only access to programs, but also to rights ensured by the Constitution.

Quasi-experiments have the potential to circumvent part of these ethical and political problems, but they can also suffer from operational implementation difficulties, which compromise the initially planned internal and external validities. Both cases, in general, imply high field costs, a well-qualified supervision structure and a good completion time (between the survey waves). Longitudinal studies integrating the records of public programs and databases can be less costly alternatives in terms of time and field operation, but they depend on the quality of available data, key variables for the physical linking of beneficiary records and effective access to data, guarded by different public agents, not always – and justifiably – willing to share citizens' personal or family information without a clear and valid reason for such.

The modeling of integrated bases with data from administrative records and public databases, created for the management of sector-based policies and programs, is an alternative to both of these modalities (Jannuzzi, 2016). Granted the quality, currency and specificity of the information recorded in these sources of data, this strategy would allow building models with many possibilities of comparative evaluation – or pseudo-randomization – of factual and counterfactual situations, "treatments" and "controls", greater or smaller interaction of differentiated programs and contexts of the populations served or of the agents operating the programs. Ex-post quasi-experimental designs could also be simulated using this methodological strategy. Moreover, and perhaps its main comparative advantage, this strategy allows, in addition, the longitudinal matching of analysis units in panels with a historical extension or periodicity that is more flexible and interesting for analyses of effects of "time or regularity of exposure to the social program".

Like any research method employed in program evaluation, there are potentials and limitations that need to be analyze for each case. But it is a fact that in these evaluation designs, public programs cannot continue to be modeled with a dummy parameterization – 0 or 1 in a signaling variable – without considering the basic characteristics of their design, particularly the intensity and time of exposure of beneficiaries to the program services and benefits, i.e., the magnitude of the "treatment dose and exposure".

The considerations systematized here about experimental and other related methods should somehow lead to a critical reflection about the stance assumed by part of the epistemic community and of evaluation practices concerning the vaunted superiority or robustness of these models in relation to other models in program evaluation. The robustness of an evaluation is not ensured by the technical sophistication, formal elegance or intrinsic theoretical qualities of the method to be employed. The robustness of an evaluation depends on the method's suitability to the problem in question, on the technical consistency with which the method is

effectively employed in the beginning, middle and end of the evaluation process, the sample, data collection and assumptions used in the analysis. The robustness of an evaluation depends on the consensus and transparency of technical choices in the face of methodological challenges that inevitably emerge in complex problems, when efforts are made to solve them according to comprehensive treatment and analysis perspectives, and not according to conveniences and inclinations of the method of preference. The robustness of an evaluation depends on the intellectual honesty with which assumptions and presuppositions about the program, its merit and results obtained through a methodological perspective are put to the test with the triangulation of other findings, obtained through other methods, subjects and interpretive perspectives.

Thus, the robustness of an evaluation is not associated with a supposed gold-standard method, but rather with a diamond-standard technical-scientific position which is enlightened, plural and vigilant, and which recognizes the contingent and limited nature of knowledge about the complex reality of public demands, collective problems, and governmental actions designed to address or mitigate them.

REFERENCES

Bamberguer, M., Segone, M., & Tateossian, F. (2016). *Evaluating the Sustainable Development Goals with a "no one left behind" lens through equity-focused and gender-responsive evaluations*. UN Women.

Batista, M., & Domingos, A. (2017). Mais que boas intenções: Técnicas quantitativas e qualitativas na avaliação de impacto de políticas públicas. *Revista Brasileira de Ciências Sociais, 32*(94), Artigo e329414. https://doi.org/10.17666/329414/2017

Cano, I. (2004). *Introdução à avaliação de programas sociais*. FGV.

Casa Civil da Presidência da República. (2018). *Avaliação de políticas públicas: Guia prático de análise ex post* (vol. 2). Casa Civil da Presidência da República.

Chianca, T. (2015). Um modelo alternativo ao estudo experimental para inferir causalidade em avaliações do impacto de projetos sociais. *Revista Brasileira de Monitoramento e Avaliação, 9,* 16-29. http://dx.doi.org/10.4322/rbma201509003

Gertler, P., Martínez, S., Premand, P., Rawlings, L. B., & Vermeesch, C. M. J. (2015). *Avaliação de impacto na prática*. Banco Mundial.

Gussi, A. F., & Oliveira, B. R. (2017). Discutindo paradigmas contra-hegemônicos de avaliação de políticas públicas. *Anais do 1. Encontro Nacional de Ensino e Pesquisa do Campo de Públicas*. UFC.

Imas, L. G. M., & Rist, R. (2009). *The road to results: Designing and conducting effective development evaluations*. World Bank.

Jaime, P. C., Vaz, A. C. N., Nilson, E. A. F., Fonseca, J. C. G., Guadagnin, S. C., Silva, S. A., Sousa, M. F., & Santos, L. M. P. (2014). Desnutrição em crianças de até 5 anos beneficiárias do programa Bolsa Família: Análise transversal e painel longitudinal de 2008 a 2012. *Cadernos de Estudos Desenvolvimento Social em Debate,* (17), 49-61.

Jannuzzi, P. de M. (2016). *Monitoramento e avaliação de programas sociais: Uma introdução aos conceitos e técnicas*. Alínea.

Jannuzzi, P. de M. (2018). Mitos do desenho quase-experimental na avaliação de programas. *Nau Social, 9*(16), 76-90. https://doi.org/10.9771/ns.v9i16.31419

Jannuzzi, P. de M., & Pinto, A. (2013). Bolsa Família e seus impactos nas condições de vida da população brasileira: Uma síntese dos principais achados da pesquisa de avaliação de impacto do Bolsa Família II. In M. Neri, & T. Campello (Orgs.), *Programa Bolsa Família: Uma década de inclusão e cidadania* (pp. 179-192). Ipea.

La Rovere, R. (2014). The consultative group on international agricultural research approach to impact evaluation on environment and natural resources management. In J. I. Uitto (Org.), *Evaluating environment in international development* (pp. 277-288). Routledge.

Leão, L. S., & Eyal, G. (August, 2016). Experiments in the wild: A historical perspective on the rise of randomized controlled trials in international development. *Comparative Historical Sociology Mini-Conference*. American Sociological Association, Seattle, USA.

Leeuw, F. L. (2010). On the contemporary history of experimental evaluations and its relevance for policy making. In O. Rieper, F. L. Leeuw, & T. Ling (Eds.), *The evidence book: concepts, generation, and use of evidence* (Comparative polyce evaluation, vol. 15, pp. 11-26). Routledge.

McShane, B. B., Gal, D., Gelman, A., Robert, C., & Tackett, J. F. (2019). Abandon statistical significance. *The American Statistician, 73*, 235-245. https://doi.org/10.1080/00031305.2018.1527253

Moral-Arce, I. (2014). *Elección del método de evaluación cuantitativa de una política pública: Buenas prácticas en América Latina y la Unión Europea*. EuroSocial.

Patton, M. Q. (1990). *Qualitative evaluation and research methods.* Sage.

Patton, M. Q. (1997). *Utilization-focused evaluation: The next century text*. Sage.

Ravallion, M. (2009). Should the randomists rule? *The Economists' Voice, 6*(2). https://doi.org/10.2202/1553-3832.1368

Rossi, P. (1987). The iron law of evaluation and other metallic rules. In J. Miller, & M. Lewis, *Research in social problems and public policy* (vol. 4, pp. 3-20). Jai Press.

Rossi, P. H., Lipsey, M. W., & Henry, G. T. (2004). *Evaluation: A systematic approach*. Sage.

Sawyer, A. G., & Peter, J. P. (1983). The significance of statistical significance tests in Marketing Research. *Journal of Marketing Research, 20*(2), 122-133. https://doi.org/10.2307/3151679

Vaitsman, J., & Paes-Sousa, R. (2009). Avaliação de programas e transparência da gestão pública. In C. Franzese, C. Anjos, D. Ferraz, F. L. Abrucio, G. N. Cheli, G. M. B. P. Melo, J. Vaistman, J. L. D. Nehmé, L. Santoni, M. G. da Silva, M. Rubio, P. Yanes, S. Nahas, P. de M. Jannuzzi, & R. Paes-Sousa, *Reflexões para Ibero-América: Avaliação de programas sociais* (pp. 11-23). Enap.

Weiss, C. I. (1998). *Evaluation research*. Prentice Hall.

Worthern, B. R., Sanders, J. R., & Fitzpatrick, J. L. (2004). *Avaliação de programas: Concepções e práticas*. Gente.