




# More Bang for Your Buck: Best-Practice Recommendations for Designing, Implementing, and Evaluating Job Creation Studies

Paloma Bernal-Turnes<sup>1,2,3</sup>  · Ricardo Ernst<sup>2</sup>

Received: 4 December 2021 / Accepted: 24 February 2023  
© The Author(s) 2023

## Abstract

This paper discusses the application of robust experimental research methodologies that help to provide a better understanding of the mechanisms of the Theory of Change, for which training programs and/or matching grants improve job creation in micro, small, and medium-sized enterprises (MSMEs and SMEs). The literature on both interventions, such as training and matching grants, recognizes methodological flaws that hamper achieving enough statistical evidence to test the aforementioned Theory of Change. A better understanding of the interventions and the mechanisms to create jobs has become critical to ensure the resurgence of the global economy after the COVID-19 pandemic and to face the threat of the upcoming industrial revolution. This paper proposes seven methodological meliorations in impact evaluation that will help to set improvements alongside the full process of a project: designing superior policies and programs, implementing projects, supporting the finer assessment of interventions, and establishing the subsequent advancement of science in testing solutions for job creation.

**Keywords** Randomized experiments · Job creation · Finance · Matching grants · Theory of Change · Impact evaluation

---

✉ Paloma Bernal-Turnes  
pb737@georgetown.edu  
<https://orcid.org/0000-0003-2553-4124>

Ricardo Ernst  
<https://orcid.org/0000-0001-7203-9535>

<sup>1</sup> Business Economics Dept., Rey Juan Carlos University, Paseo de los Artilleros s/n, 28032 Madrid, Spain

<sup>2</sup> Global Logistics Research Program, McDonough School of Business, Georgetown University, 37th and O St. NW, Washington, DC 20057, USA

<sup>3</sup> The World Bank, 2121 Pennsylvania Ave. NW, Washington, DC 20433, USA

## Introduction

Impact evaluation is the tool to support evidence-based policy making because it serves as a foundation for managing projects with a better assessment of the performance of development programs and policies. The evidence derived from a strong methodology in impact evaluation is crucial when challenging and developing new theories, thus creating knowledge about what, how, and in which extension a certain intervention contributes to a specific goal. Indeed, the attribution of what and how it can generate the desired outcomes (which is the definition of the Theory of Change) is a question of paramount significance in both economic and the entire social science; this is due to the fact that the peerless contribution of an impact evaluation is to determine and estimate causal relationships between policies and outcomes. Thus, a greater knowledge of causal effects could help to solve the crucial problem of job destruction after the pandemic of COVID-19, if we could only find out what, how, and in which extension a specific intervention creates jobs.

The need for improving our knowledge of causal effects has been claimed by many researchers in social science (Shadish et al., 2002; Uy et al., 2010; Miller and Tsang, 2011; Allen et al., 2014; Podsakoff et al., 2014). The required methodologies for understanding the direction and nature of causal relationships (Spector, 1981; Grant and Wall, 2009; Aguinis and Bradley, 2014) are the true field of experimental methods which consist of randomized and quasi-randomized experiments (the latter are also called “quasi-experiments”). We are warned that methodological advancement tends to disseminate slowly (Aguinis et al., 2009), and it has been argued that management scholars do not know how to conduct field experiments (Borsboom et al., 2009; Highhouse, 2009). Thus, building a robust methodology in impact evaluation will provide us with the evidence that an action may distinctly, unambiguously, and almost unerringly cause a certain outcome of interest and the means by which the results occur (Lopez-Acevedo and Tan, 2011; Gertler et al., 2016). However, experiments in social science are scarce, as can be seen in the literature revisions in top journals (Podsakoff and Dalton, 1987; Scandura and Williams, 2000; Austin et al., 2002; Grant and Wall, 2009; Cravo and Piza, 2016; Buba et al. 2020; Dvouletý et al., 2021). Additionally, even when research is based on experiments, the analysis of concrete and attributable findings from impact evaluation on a specific topic is not always an easy task for practitioners, donors, or researchers, due to the lack of comparability of the studies.

The complexity to compare outcomes in impact evaluations in social science comes from a wide variety of sources. First, one of the difficulties to extract valid information from impact evaluation cases comes from the different perspectives that stakeholders could have about any project. Donors and policymakers are more likely driven by the outcome of a project, the “what to achieve,” such as job creation, because they are involved in providing solutions for a certain problem, which could be the reduction of the high unemployment rate. In contrast, other stakeholders involved in impact evaluation projects, such as

practitioners and economists, are more likely to focus their attention on tools and inputs to achieve those outcomes, the “how to achieve it.” Thus, the provision of the service becomes the essence of the monitoring mechanisms. Continuing with this example, the “how” to achieve job creation could be done by a wide variety of policies and services, such as the provision of the application of new technology, basic education, matching grants, or capacity building. We argue that it is important to reconcile the different perspectives that link outcomes and inputs, theory and practice, and policies and facts, in the belief that they are interrelated. We hope that this paper encourages scientists and practitioners to build bridges between problems and solutions for which impact evaluation projects are designed.

Second, when research is based on experiments, the analysis of concrete and attributable findings from impact evaluation requires transparency in reporting results, as this will help assemble evidence from programs. Meta-analysis of research literature, longitudinal studies, and Bayesian analysis facilitate the understanding of the generalization of causal effects with different participants, time, treatment variations, settings, and research methodologies (Hedges and Olkin, 1980, 1985; Hedges, 1987; Howard et al., 2000; West and Thoemmes, 2010). We highlight that there is not yet solid evidence that the creation of jobs in SMEs is attributable to matching grants (Hristova and Coste, 2016; Piza et al., 2016; Cravo and Piza, 2016). Indeed, the revision of the literature on the effects of the provision of finance with matching grants on job creation provides a too wide range of conclusions that go from inconclusive results (Tan and Lopez-Acevedo, 2005; Bruhn and Love, 2012; McKenzie et al., 2015), negative effects (Rijkers et al., 2010; Karlan et al., 2015; Fiala, 2018), inaccurate effects (Lopez-Acevedo and Tinajero, 2010), and non-effects (Bruhn et al., 2012). More transparency when reporting will avoid conflicting results.

This paper exposes a set of methodological improvements that define a rational process to increase accuracy in impact evaluation, followed by the advancement of science. These methodological improvements are the methodological dreams that we would like to see come true in the years to come, with the hope to get conclusive findings and advance the practical and theoretical knowledge about job creation in SMEs. By reviewing the literature about the intervention training and matching grants, we have identified some relevant aspects that are not sufficiently addressed to properly test the Theory of Change. The following are the impact evaluation improvements we dream of for the methodological research flaws:

- 1) the time frame to test theories;
- 2) the sample size;
- 3) the effect size and power;
- 4) the descriptive analysis;
- 5) the budget allocation;
- 6) the generality of results; and
- 7) the outcomes.

## Advancement in Theoretical Progress

Acknowledging that the advancement of science relies on the accumulation of knowledge, “social science” (such as Economics or Business Management in the topic of job creation) versus “hard science” (such as Medicine, Physics, or Astronomy) faces several threats in this empirical comparison (Borsboom et al., 2009; Highhouse, 2009). The acid test for the theoretical advancement is the replicability of results (Hedges, 1987), but specific aspects in social science obstruct theoretical progress (Pfeffer, 1993) and the consensus on scientific paradigms (Davis, 2010). In Economics and Business Management stand out the lack of standardized tools and unstructured methodological procedures (Ketchen et al., 2008; Edwards, 2008; Hitt et al., 2004; Aguinis and Edwards, 2014) which deter the performance of experiments as it is similarly done in hard science for ensuring, unerringly, the attribution of results (Cravo and Piza, 2016). More specifically, in the realm of job creation in SMEs within the Theory of Change, the comparison of social science experiments is difficult because there is ambiguity and imprecision when standardizing core concepts and tool definitions (Cravo and Piza, 2016; Grimm and Paffhausen, 2015), such as the concept of SME; not all environments and legal frameworks have accepted the consensus on the threshold of 250 employees (Ayyagari et al., 2007; Cravo et al., 2012), which limits consistency in the results and the development of theories.

Replicability in social science also requires isolating the effects of pervasive interactions (Cronbach 1975), historical and cultural influences (Gergen, 1973, 1982), unstandardized measurements (Kruskal, 1978), and blurry indirect and moderated effects (King et al., 2012; Bernal and Ernst, 2015, 2016; Eden et al., 2015) to achieve more precise and more unerringly attributable outcomes for SME interventions.

Additionally, researchers suggest that the advancement of science needs more than the rejection of the null hypothesis, even if the statistical power is acceptable (Meehl, 1978; Bezeau and Graves, 2001). Solving big economic and social dilemmas of our era, such as job creation, requires different data collection techniques from different sources of data, multi-method statistical analysis, and novel uses of technologies (Gregorie et al., 2010). The integration of new technologies in true field experiments is opening new opportunities to test theories and measure concepts of social science with more accuracy (Aguinis and Lawal, 2012; Hsu et al., 2017). Economics should rely more on complementary formats for doing research, integrating different analysis techniques such as descriptive analysis, graphical methods, geospatial analysis, and meta- and Bayesian analysis.

## Time Frame to Test Outcomes

The first methodological melioration in impact evaluation is linked to the length of the experiment; it should be based on the nature of the program and always be embedded in the literature on the specific intervention and the outcome. It has been suggested that the length of the program is one of the main threats of previous

studies to sufficiently capture the impacts searched (Cravo and Piza, 2016). The right length of the program depends on the magnitude of the expected outcomes, the trend of which could follow a linear or a non-linear pattern. Indeed, the evaluation of the full effect of some programs, such as training (Mano et al., 2012) or technology diffusion (Hall and Khan, 2003), requires both a prudent time to elapse between intervention and outcomes (because short-run effects are limited) and several rounds of follow-up surveys that build panel data to observe the efficiency of a policy. Multiple waves of measurement will allow us to shape the curve-fitting relationship between intervention and outcome. For example, the learning curve of new skills or the productivity curve of technology adoption is impossible to observe with only two-time measurements (pre- and post-treatment), and it could be even worse, if the time that elapses between the baseline and the end line of the program is too short, the outcomes should be imperceptible, and therefore, not statistically significant.

Our methodological dream for the coming years is to see more publications in impact evaluations that test the results of training and/or matching grants in several waves to capture curve effects (Mano et al., 2012; Histrova and Coste, 2016). More specifically, the desirable timeframe to observe results for matching grant projects to create jobs is approximately 2 years (Rijkers et al., 2010; McKenzie et al., 2015), while training interventions have a shorter-run impact on performance and productivity (McKenzie et al., 2015); however, it could be imperceptible in the short-run. Additionally, it is important to contextualize the stage of entrepreneurial experience when inferring causality (Hsu et al., 2017). It has been tested that entrepreneurship activity unfolds over time and incurs differently in the outcomes (Shane, 2003; Shane and Venkataraman, 2000; Haltiwanger et al., 2013).

## Sample Size

The second recommendation for doing more robust impact evaluations is the need to perform more accurate calculations of a minimum sample size, which is needed to test the impact of any intervention. The size of the sample has a tremendous influence on research costs, but it also affects the quality of the study. Unfortunately, it is possible to read that sample sizes that are too small are the reason for having inconclusive results in some project interventions (Cravo and Piza, 2016). Too small a sample size can lead to the paradox of obtaining very different means in the treatment and the control, while that difference is not statistically significant. In other words, differences in means between treatment and control groups are happening by coincidence, instead of happening in every sample extracted from the population. The paradoxical failure of using a sample size that is not large enough drives us to either useless interventions to achieve desired outcomes or, what could be even worse, to reach real outcomes that are not observable from an effective intervention, whatever could be the significance level of acceptance of the hypothesis (Khandker et al., 2010; Gertler et al., 2016). Acknowledging the problem caused by small sample sizes in specific impact evaluations, conclusions such as a much higher survival rate of those MSMEs whose entrepreneurs receive training (Mano et al., 2012), or a much higher increase in the business performance of those MSMEs that receive

consultancy services, could not be taken into consideration for being statistically insignificant (Bruhn et al., 2010).

The sample size is also linked with the significance level (called alpha), or the Type I error rate, which is the probability of rejecting  $H_0$  (null hypothesis) when it is actually true. The smaller the Type I error rate, the larger the sample required for the same power. Power in statistics identifies how reliable the results obtained are, which is linked to the Type I error rate. In other words, the larger the sample size, the higher the Type I error.

The calculation of the sample size is linked with the calculation of the statistical power, which is higher as the size of the treatment and control group are equal (Cohen, 1994; Khandker et al., 2010; Gertler et al., 2016). Besides the sample size calculation to achieve a specified standard error and a certain probability of statistical significance (Gelman and Hill, 2006), the sample size must be big enough to easily identify similar means and standard deviations in the control and treatment groups in the baseline. The effect size is the metric that calculates the distance between the control and treatment groups in means and standard deviations. We should be able to detect the smallest effect size possible because it will help to directly attribute the outcomes of the intervention in the endline.

Impact evaluations, when well-constructed, build two subsets of the sample with the following criteria: (1) the treatment group must be a reliable and a fair representation of the full population, (2) the control group must be the mirror of the treatment group in terms of its main characteristics, and (3) replications of the study will bring similar results. These requirements to build the treatment and the control group (analysis ex-ante) and rigorous empirical methods to analyze the causal effects (analysis ex-post) will allow to determine the outcomes that are uniquely, distinctly, and directly attributable to the matching grants (Khandker et al., 2010; Gertler et al., 2016). Thus, SMEs after being randomly reached out with open access communication campaigns to apply for the financial support, which ensures equal opportunities to participate in the intervention, must be assigned the treatment through randomization.

Consequently, a comparison group will be constructed from the applicant pool of those firms that did not receive any grant, but entities are similar to the first sample on observed pre-intervention characteristics. Linked to the decision of the method to be used to build the comparison group and estimated the counterfactual, we need to calculate the sample size.

Our methodological dream for the years to come is to see more publications on impact evaluations. These can provide sample size calculations for the building of the control and treatment groups and for performing the main analysis of the experiment tailored by the type of statistical methodology applied and the number of variables (and estimations) considered.

## Power and Effect Size

Most literature in impact evaluation is focused on reporting the statistical significance of the validity of a null hypothesis, as it is reported in any other scientific area

of knowledge. However, power analysis very often is underreported in social science or scarcely done with a transparent approach (McKenzie et al., 2015).

The power analysis provides a unique piece of information in impact evaluations: The implied probability of making an error of estimation. The power of a statistical test of a null hypothesis is the probability of making the correct decision, that is, the probability of being right in the rejection of the null hypothesis if the alternative hypothesis is true (Cohen, 1988; Cohn and Becker, 2003). When power is low, the more likely it is to make a Type II error, which is the probability of rejecting the null hypothesis when the alternative hypothesis is true. In the pre-treatment phase of any experiment, even with significantly similar means in the treatment and control groups, with low power, there is a higher risk to assume that both groups have similar means when it is false. In other words, there is a high risk to get inconclusive results from an experiment in which the control group differs significantly from the treatment group. Under this scenario, it is important to realize that with Type II errors, it will be quite unlikely to determine that the outcomes are uniquely, unerringly, and fully attributable to the intervention executed. Then, the Theory of Change will be wrongly texted in the post-intervention analysis under high Type II errors in the pre-treatment analysis.

An example to illustrate that low power could lead us to a wrong conclusion is the case in which the treatment group has a higher job creation rate than the control group, but instead, the test of means differences (effect size) in the pre-treatment stage leads us to think that both groups are significantly similar. In this case, if the power of the test is very low, then, we will be more likely to wrongly assume that the intervention to facilitate access to finance (executed only in the treatment group) is undoubtedly the reason for having a significantly higher job creation rate than the control group, which is false. Only with the power test can we check the probability of being wrong in the conclusion. When we increase power, we reduce Type II error, and we are more likely able to say that our findings are robust, because firms under equal conditions (the job creation means are significantly equal in a pre-treatment test) create more jobs when they have access to the finance intervention (during the post-treatment test). The key point here is that, under low-powered testing, effects statistically significant tend to vary greatly from different samples, producing patterns of apparent contradictions in the published literature (Maxwell, 2004; Cohn and Becker, 2003), reducing theoretical precision that impedes the generalization of policies and programs. In other words, seriously underpowered impact evaluations are useless, while, in turn, the increase in the statistical power builds coherence in the literature and advances scientific knowledge (Maxwell, 2004).

Literature shows the persistence of lack of distinction in the power analysis done in each experiment (Cohen, 1994; Saris and Satorra, 1993; Satorra and Saris, 1985). Indeed, any single study should include multiple power analyses (Maxwell, 2004) such as the incremental explanatory power analysis (Biscotti and D'Amico, 2019). First, the experimental design will require testing the power in building the sample size to determine the number of subjects needed in the study to detect an effect of a given size (Cohen, 1994). Additionally, power analysis must be done to design the main statistical analysis tailored to the research

methodology applied: the larger the number of explanatory variables, the larger the sample size is required (Maxwell, 2004).

Power analysis and sample size calculations could be conducted using Cohen's tables (Cohen, 1994) or software such as G\*Power (Faul et al., 2009), SAMPLE POWER (SPSS, 2017), and R (Green and MacLeod, 2016). Calculations of statistical power depend on the alpha significance, the sample size, and the effect size (Bezeau and Graves, 2001). The most common way to solve underpowered analysis is by increasing the sample size, which easily raises the cost of the experiment. However, other three less costly actions allow for the increase of statistical power in the post-treatment tests. First, the use of more advanced applied methodological techniques, such as multilevel analysis (Kozlowski and Klein, 2000) and the addition of covariates (Satorra and Saris, 1985; Judd and McClelland, 1989; Maxwell et al., 2018). Second, the formulation of simple null hypotheses, rather than formulating complex null hypotheses (McClelland, 1997). Third, in case we wish to test a dichotomic outcome, the use of a more efficient allocation of observations that maximize the variance with half of the sample in the two extreme values ( $\frac{1}{2}$  0 0  $\frac{1}{2}$ ), instead of using a standard normally distributed across the mean ( $\frac{1}{4}$   $\frac{1}{4}$   $\frac{1}{4}$   $\frac{1}{4}$ ) (Mead, 1988; Atkinson and Donev, 1992; McClelland, 1997).

Additionally, in the power analysis, effect size should be reported too. Effect size measures the distance between the treatment and the comparison group (Bezeau and Graves, 2001). If the research is able to detect small effect sizes of a treatment, it will lead to a better detection of the causal effects between interventions and outcomes. Cohen's (1988) conventional definitions of small, medium, and large effect sizes for each statistic measure are usually the most commonly used tool (Mone et al., 1996).

Cohen (1988, 1992), Hunter and Schmidt (1990), Lykken (1968), Rosenthal (1991), and Mone et al. (1996) highlighted two advantages of reporting and evaluating effect size in research. First, the effect size reports the magnitude of the phenomenon in the population (Mone et al., 1996), and then, the comparability of studies. Second, reporting effect sizes increases the replicability of research streams and the comparability of studies with meta- and Bayesian analysis (Maddock and Rossi, 2001; Hedges and Olkin, 1980, 1985). Notice that the problem of comparing studies with unequal effect-sized groups is the invalid comparison of chi-squares that have different degrees of freedom (Hedges, 1987).

There are, at least, three ways to increase effect size, besides increasing sample size. First, the use of more advanced methodological techniques, such as multilevel analysis (Kozlowski and Klein, 2000). Second, the use of more reliable measurements using the formulae suggested by Schmidt et al. (1976) or Schmitt and Klimoski (1991) for increasing the validity of estimations to increase the statistical power and reduce the need for larger samples (Sawyer and Ball, 1981; Schmidt et al., 1976; Sutcliffe, 1980). Last of all, the use of statistical methods of Cascio and Zedeck (1983) for reducing the alpha level and then increasing the power.

Our methodological dream for the years to come is to see more publications in impact evaluations that provide and report the analysis of power and effect size. It is a tenet of good practice if both calculations are made while designing the impact evaluation. We encourage researchers to use advanced statistical methodologies, besides sample size, to increase statistical power and the capacity to detect small effect sizes.



## Descriptive Analysis

Impact evaluation is not an exception in Economics for not carrying out a deep descriptive analysis of what the data shows in both the treatment and control groups before conducting inferential statistics. Traditionally, studies that use impact evaluation select the control group in terms of similarity of the central tendency measures, mainly mean or median, with the treatment group. The assumption of normality for the treatment and control group should not be assumed, and instead, it must be analyzed in the baseline, especially for experiments with small sample sizes. Once the normal distribution is tested in the treatment and control group, the larger the standard deviation in the control group, the higher the probability of making error type II in our conclusions. Then, the analysis of the standard deviation is critical for increasing internal validity and power, besides a bigger sample size. Our methodological dream about the descriptive analysis is to incorporate the analysis of the measure of variability with the standard deviation in the control group. We hope to see more publications with experimental and quasi-experimental designs that provide complete descriptive analysis and test the similarity of means and standard deviations in the treatment and control groups in the pre-treatment phase.

## Incorporating Realistic Challenges During Implementation

Previous methodological studies about impact evaluation are not naive in assuming changes that could arise during the implementation, and indeed, they reveal that evaluation designs quite often are not implemented as initially stated (Gertler et al., 2016). This section is created on the belief that some research methodological improvements for succeeding in inferring causation require close cooperation among practitioners and researchers. Said cooperation could help to

- eradicate biased samples,
- identify context-specific phenomena,
- alert of events that disrupt outcomes (such as changes in legal frameworks),
- identify and collect cofounder variables, and
- build a common understanding that enriches the perspective of the research.

The combination of experience on the ground, the strong cooperation with local stakeholders and donors, and the advice from researchers and economists could help to identify the right research design and execution (Grant and Wall, 2009; Rynes and Bartunek, 2017). Indeed, impact evaluation not only advances knowledge but also improves project implementation itself, as it helps to allocate resources and to increase the accountability of the project (Legovini et al., 2015).

## Budget Allocation

Monitoring and evaluation systems allow for the implementation of programs and interventions with transparency and accountability for the sake of an effective budget management. When more than one intervention is executed at the same time, researchers should design and report the scientific findings with a clear distinction of the following three aspects: (a) samples (from beneficiaries and control groups), (b) interventions, and (c) budgets, to study the effectiveness of each intervention with a proper and precise estimation of the causal effect to achieve a specific outcome (Lopez-Acevedo and Tinajero, 2010).

Clear accountability of interventions allows showing the results achieved in measurable outcomes. These could be translated into a convertible currency, personnel, or time length, which provides a better understanding of the findings and facilitates the attraction of investors by explaining, for instance, the dollars needed to create each job, the personnel needed to provide training to increase a certain amount of revenue, or the number of months needed to create each job position. This improvement in budgeting interventions helps researchers to generalize the study. Additionally, researchers shed light, not only on the effect size (in our examples in dollars or months), but on the importance of the “cause size” in comparing experiments (Highhouse, 2009).

Our methodological dream for the years to come is to see more publications on impact evaluations that facilitate independent budget allocation of each intervention and samples separately and, additionally, specific budget allocation for each interaction of programs when they exist.

## Generalization of Results

The advancement in knowledge in answering a cause-effect question requires the generalization of results achieving broader effectiveness and scalability under two different angles: on the one hand, by testing alternative programs to achieve the same outcomes, and on the other, by testing a causal effect with the Theory of Change applied in the distinctiveness of research settings (Cook and Campbell, 1979).

The first angle of analysis consists of testing the effectiveness of a series of alternative programs in a particular setting, timeframe, and population, which contributes to the generalization of results, as it validates the status of the theory when, separately, different interventions are applied in the same scenario to achieve the same outcome, such as creating jobs either by providing matching grants (Hristova and Coste, 2016; Piza et al., 2016; Cravo and Piza, 2016) or by facilitating access to external markets (Rossignol and Salmon, 2016). Then, the body of evidence became more robust concerning the bundle of benefits, threats, and spillovers when reaching a specific goal under different interventions in the same environment (Highhouse, 2009; Rijkers et al., 2010). This approach not only advances science but also brings very valuable information to policymakers, because it provides information about the policy that is more adequate to solve a problem.

The other angle to generalize results consists of testing the properties of a certain theory that should be applied in multiple settings or audiences, such as testing a specific mechanism to create jobs, in both peaceful versus conflicted and violent environments. Under this angle, the alternative explanations of the effects have been isolated, and the remaining attributable effects of the intervention are tested separately in both scenarios. Only then, the generalization of results could be possible, because the analysis provides the information on the effectiveness of a certain intervention with and without the different circumstances that affect the scenario, recognizing the minimum and maximum effect of an intervention: either the environment is peaceful or conflicted and violent. Research shows that the mechanism to create jobs in conflict and violent environments requires the analysis of the effects of uncertainty on entrepreneurs' decisions for taking risks (Knight, 1921; McMullen and Shepherd, 2006; Ralston, 2014; Mallet and Slater, 2016); in other words, it should require control for uncertainty to do the variables consistent across conditions (Hsu et al., 2014, 2017; Ashta et al., 2021). The studies made by McKelvie et al. (2011), Koudstaal et al. (2015), and Holm et al. (2013) are experiments that include the analysis of entrepreneurs' decisions and actions process under conditions of uncertainty. Indeed, job creation in conflict and violent environments is a way of building social cohesion, allowing the transformation of informal into formal businesses, and improving the inclusivity of ex-combatants and potential insurgents (Ralston, 2014).

Under both angles of building knowledge, the reduction of the sources of bias increases the external validity and then increases the generalizability of the results. The methodological ameliorations that reduce the sources of bias can be achieved by correctly identifying and clearly explaining the following items: (1) the selection of the population of interest to the research question asked, (2) the attributes of the context that influence the sample or subsets of the sample, (3) the active or passive nature of the individuals analyzed (Hsu et al., 2017), (4) the research question embedded in the field of science, and (5) the adoption of the statistical methodology tailored to the research question and the program's operational characteristics (Khandker et al., 2010; Gertler et al., 2016).

Additionally, from both angles, the advancement of science in job creation could be trapped in not having enough proof to distinguish between null findings that result from low power (Cohn and Becker, 2003) and null findings that reflect a genuine absence of effect size that results from the wide length of the confident interval (Howard et al., 2000). This dilemma is not yet solved in experiments in SMEs, since its literature scarcely reports confidence intervals (McKenzie et al., 2016) and power analysis as has been discussed before. Besides the contribution to solving this dilemma about null findings, reporting confidence intervals around the mean provides three main additional advantages. First, confidence intervals provide information about the precision of the estimate because it implies the value of a hypothesis test, which is the zero value within the interval. Thereby, the precise estimation of the null hypothesis of no difference could not be rejected in a tiny degree of error, generally stated at an alpha level of .05 (McCallum et al., 1996; Howard et al., 2000). Second, confidence intervals provide further information beyond "yes-or-no" outcomes (the yes or no non-zero population effects), as the smaller the confidence

intervals, the higher the precision and the better the estimation of effect size (Cohen, 1994; Cohn and Becker, 2003). Finally, confidence intervals facilitate the theoretical interpretation of its central point (Bezeau and Graves, 2001).

Literature also reflects the concern that the reliance on tests of statistical significance contributes to the poorer theoretical and empirical cumulativeness of knowledge in social science that hampers the generalization of results (Meehl, 1978; Hedges, 1987). Meta-analysis of studies and Bayesian analysis, however, can increase the likelihood of detecting population effects (Hedges and Olkin, 1985; Hedges, 1987; Hunter and Schmidt, 1990; Maddock and Rossi, 2001; Cohn and Becker, 2003; Aguinis and Edwards, 2014), and they focus on the magnitude of a treatment effect, such as the magnitude of jobs creation through matching grants.

Our methodological dream for finding an answer to the questions about job-creating interventions in an unbiased way is to encourage authors to state the findings in terms of the magnitude of the effects with their confidence intervals in order to better answer this theoretical conundrum (Cohen, 1994; Aguinis and Edwards, 2014). Then, the comparison and assembly of evidence through confidence intervals is linked with the use of meta-analysis, Bayesian analysis, and longitudinal techniques and could solve the problems created by low statistical power in individual studies (Hunter and Schmidt, 1990; Cohn and Becker, 2003).

## Outcomes to Test the Theory of Change

Randomized experiments are not always the gold standard for research design in social science for the advancement in the knowledge of the Theory of Change (Campbell and Stanley, 1966; Cook and Campbell, 1979; Grant and Wall, 2009). Quasi-experiments are also fundamental for building and generalizing strong theories in social science (Dubin, 1976; Whetten 1989; Grant and Wall, 2009). Indeed, quasi-experiments ensure the rigorous construction of boundary conditions under a certain treatment, which is more or less likely to exert a particular pattern of effects (Johns, 2006). Thus, it is the ideal analysis for interventions that affect certain groups differently within the sample (Khandker et al., 2010; Gertler et al., 2016). Additionally, quasi-experiments are for unethical problems to randomize the treatment and the control group (Khandker et al., 2010; Gertler et al., 2016). For these two reasons, quasi-experiments require judicious research choices and rigorous methodologies that expand internal and external validity.

Quasi-experiments have traditionally been seen as a silver medal for testing causal effects (King et al., 2012), and the gold one was designated for randomized experiments (Shadish et al., 2002; King et al., 2012). Other authors, however, defend the advantages of quasi-experiments over randomized experiments in terms of implementation, validity, and testing (Campbell and Stanley, 1966; Cook and Campbell, 1979). Richer discussions have also emerged about experiments (Highhouse, 2009; Bullock et al., 2010; Aguinis and Lawal, 2012; King et al., 2012; Eden, 2017). Indeed, generalization in social science requires careful attention by cofounders (McKelvie et al., 2011; Holm et al., 2013; Hsu et al., 2014,

2017; Koudstaal et al., 2015), sampling stimuli, and strengthening manipulations (Highhouse, 2009) to undoubtedly understand causal effects.

Estimating the impact of the treatment in quasi-experiments depends on constructing a valid counterfactual group that parallels the SME beneficiary group in all respects except for participation in the intervention under evaluation. For this purpose, Propensity Score Matching (PSM) will be used to create statistically equivalent counterfactuals to the treatment group. The evaluation question, expressed in the following expression, can be simplified as estimating the average treatment effect by taking the difference between the expected outcomes of the treatment and comparison groups:

$$ATE = [E(Y_i(1)|T = 1)] - [E(Y_i(0)|T = 0)]$$

where “ATE” is the average treatment effect and “ $Y_i$ ” is the outcome of the  $i$ th SME unit.

The outcomes for the treated and comparison SME units are

$$Y_i(1) = Y_i(T = 1) \text{ for the matching grant treatment } T \text{ (treated)}$$

$$Y_i(0) = Y_i(T = 0) \text{ for the matching grant treatment } T \text{ (comparison)}$$

PSM addresses the problem of a missing counterfactual: The fundamental idea of the PSM approach is that for each unit in the treatment group and in the pool of non-selected firms, the probability of treatment (propensity score) is computed based on observed characteristics. Background covariates of selection (e.g., firm size, age, number of employees, sector) into treatment are converted into this single scalar propensity score, thereby reducing multidimensionality. The score, ranging from 0 to 1, is the SMEs probability of receiving treatment conditional on observed covariates. This quasi-experimental approach ensures that the average characteristics of the treatment and comparison groups are similar, which is a necessary condition to obtain unbiased estimates. The impact of grants on beneficiaries can be estimated by comparing the average outcomes of a treatment group and the average outcomes of a statistically matched subgroup of firms, the match being based on observed characteristics available in the data at hand.

Applying a matching evaluation design would translate into the following broad steps:

- 1) Estimate the propensity score using either the probit or logit model.  $P[x] = P[T = 1 | x]$ , i.e., the probability of receiving the treatment [ $T = 1$ ] given a set of observed characteristics.
- 2) Choose an appropriate matching method to match the estimated propensity scores of treated SME units to untreated SME units—Methods such as nearest neighbor, radius, stratification, kernel, caliper, and others can be used. We propose using the 2:1 nearest neighbor technique, based on the principle of minimizing the absolute difference between the estimated propensity scores for the control and treatment groups.

- 3) Assess the quality of the matching by checking for common support and balance, and as a result, restricting the sample to units for which common support appears in the propensity score distribution.
- 4) For each treatment unit, locate a subgroup of comparison group units that have similar propensity scores.
- 5) Compare the outcomes for the treatment units and their matched comparison units. The difference in average outcomes for these two subgroups is the measure of the impact that can be attributed to the program for that particular treated observation
- 6) The mean of these individual impacts yields the estimated average treatment effect or ATE (i.e., difference in outcomes between the participants and matched non-recipients).

Our methodological dream is to see more publications that coherently build the realm of knowledge with either quasi- and randomized experiments to build the Theory of Change that can explain the causal effect of job creation or quasi-experiments to understand the circumstances, contexts, and groups that exert differently in terms of job creation.

## Conclusions

The main originality of this paper is to identify the main methodological meliorations that support evidence-based policy making for the creation of jobs. The surge in the demand for a better assessment of the performance of job creation programs and policies has become critical in providing prosperity. This paper should drive practitioners and researchers to benefit from methodological improvements for inferring causation in financial support to SMEs to create jobs. The revision of previous studies shows evidence that firm-level experiments focused on the impacts of interventions on job creation are complex to infer positive direct effects (Buba et al., 2020; Dvouletý et al., 2021). More often than desired, the provision of financial support to SMEs provoked unseeking results on job destruction among the beneficiaries of the interventions (Fiala, 2018; Karlan et al., 2015; Rijkers et al., 2010). In this regard, although implementation and monitoring are at the heart of evidence-based policy making, the impact evaluation should also use a core set of statistical tools to ensure that the outcomes on job creation are uniquely, unerringly, and fully attributable to the intervention executed. Our hope is that by linking implementation best practice recommendations to the methodological meliorations in impact evaluation, we will be able to catalyze further research in the area and, ultimately, to support more efficient policies that bring prosperity through job creation.

## Theoretical Implications

Our study identifies methodological meliorations that ensure robust statistical evidence and accurate assessment of the true impact of financial support interventions in SMEs. Most of the methodological improvements suggested in this

paper obey the concern about how fundamental pre-intervention analysis, such as sampling, matching, and time framework, is in causation (Rubin, 1974, 2007, 2008; Cook and Steiner, 2010). This thought was very nicely stated by Campbell (1969a, 1969b), Rubin (2007, 2008), and Cook and Steiner (2010), who said that it is not possible to put right with statistics what has been wrong by design.

Our study suggests that the decision of applying randomized versus quasi-experiments has some trade-offs. Randomization maximizes internal validity but, in contrast, undermines external validity, while quasi-experiments in many organizational settings could provide superior external validity with good levels of internal validity (Grant and Wall, 2009; Campbell and Stanley, 1966).

This paper shows that, apart from increasing sample size, the reduction of the standard deviation in the control group brings higher power, as well as the subsequent reduction of Type II error in the analysis. But, in contrast, just by increasing the sample size, it boosts the Type I error. For this reason, power at the level of .8 is acceptable in experiments to solve this trade-off and to more confidently detect and reject false null hypotheses (Martínez, 2022). Only in two situations are unpowered studies justified (Halpem et al., 2002): (1) in interventions that are aimed to solve situations that affect either odd cases or a very limited number of individuals; (2) in early-phase trials for defining better ulterior purposes of an intervention.

In our methodological dreams, impact evaluations always report effect sizes (besides power analysis too). For the determination of the control and treatment group in the matching pair analysis, a small effect size is recommended, while for the main statistical analysis that tests the treatment effects on outcomes, a medium effect size seems an appropriate standard to do the power analyses (Cohen, 1962, Bezeau and Graves, 2001). However, it is relevant to highlight that important findings were detected when an effect size was really small in the main statistical analysis of the experiment, as it happened with the finding that headache pills reduce heart attacks. This experiment was done with a very large sample size of 22,000 individuals and a very small effect size (.0022). Social science literature, and specifically economic impact evaluation studies, should reveal power analysis more often and link and reveal a certain level of power (at least .8) with the smallest effect size possible (Mone et al., 1996).

Similar information on effect sizes is provided by confidence intervals (Cohen, 1994; Bezeau and Graves, 2001). Confidence intervals reveal the status of the null hypotheses and the non-nil null hypotheses and facilitate the generalization of knowledge because it allows for the comparison of results. Based on the advantages of calculating confidence intervals, why do authors not report them? Findings suggested that confidence intervals are the thermometer of imprecise findings (Cohen, 1994; Howard et al., 2000). Any underpowered research is more likely to incorrectly accept false null hypotheses that contain erroneous conclusions about the hypotheses tested in impact evaluation (Mone et al., 1996; Tversky and Kahneman, 1971).

Erroneous conclusions hamper the advancement of science because it provides conflicting results as well as conclusions (Smith, 1977; Grant and Wall, 2009). The statistical explanation for conflicting results in impact evaluation is named Type III

error. Type III errors occur when the null hypothesis is false and is rejected for being the direction of the true population contrary to the direction of the observed difference (Kaiser, 1960; Leventhal and Huynn, 1996; Highhouse, 2009). The problem is that it is difficult to detect Type III errors in social science. It is possible to reduce Type III errors when conflicting impacts are tested if sample stimuli occur. A recommendation for avoiding conflicting results in social science experiments is to control the mechanisms in which control and treatment groups could share information or attitudes that affect the outcomes. Conflicting results could appear in non-linear causal effects, for which the analysis of mediation and moderation effects could provide a wider picture of the intensity, and even the direction, of the effects that a treatment exerts on the outcomes (Bullock et al., 2010; King et al., 2012; Eden et al., 2015). Another solution to avoid selection bias and get the true impact of financial support to SMEs is to test the robustness of results applying Rosenbaum's (2002) bounding approach (Alemu and Ganewo, 2022).

But also richer discussions have emerged about experiments (Highhouse, 2009; Bullock et al., 2010; Aguinis and Lawal, 2012; King et al., 2012; Eden, 2017). Indeed, generalization in social science requires more careful attention to cofounders (McKelvie et al., 2011; Holm et al., 2013; Hsu et al., 2014, 2017; Koudstaal et al., 2015) and sampling stimuli (Highhouse, 2009) to undoubtedly understand causal effects.

A clear budget allocation is a "should" that, besides the advantages of accountability, helps researchers to show findings in a tangible and appealing way for the comparison of studies.

One of our methodological dreams also reflects a concern about the timing of the experiment, in which the length of the project should be adapted to the environmental challenges. Experiments also allow testing the Theory of Change with temporal progression (Lopez-Acevedo and Tinajero, 2010). Time series analysis provides stronger empirical evidence and provides the course, the strength, and the direction of the outcomes.

In summary, if robust research methodologies are as important as we believe it to be, research that yields insight into the mechanisms behind their development and the strategic choices on which they rest could make an important scientist contributions on a wide variety of topics.

## Managerial Implications

Impact evaluations are needed to inform policymakers on a range of decisions, from curtailing inefficient programs, to scaling up interventions that work, to adjusting program benefits, to selecting among various program alternatives. Business support interventions focused on SMEs are crucial since it has been tested that SMEs generate the majority of employment in developed and developing countries (Ayyagari et al., 2011). Unfortunately, SMEs encounter astounding low productivity performance and overcome barriers to grow (Mead and Liedholm, 1998, Alemu and Ganewo, 2022). These two aspects support high-priority policies that target SMEs especially in environments that face several constraints including no access



to finance, shortage of equipment, low productivity, outdated technology, and lack of skilled labor forces. Matching grants sounds that could address these constraints in situations where formal financial institutions are not willing to take any exposure beyond very basic financial services such as deposit collection, payments, and remittances. The market failure is thus evident from the credit crunch aggravated by the post-COVID scenario, especially in less developed economies and fragile and conflict environments.

Beside these market imperfections to allocate financial resources, the analysis of results shows that there are two managerial aspects underlying suboptimal allocation of inputs that have implications to design more effective policies: (1) behavioral biases—such as misperception of returns associated with a given business practice, lack of motivation to adopt better production process (Gibbons and Henderson, 2012), and cultural barriers to access to formal financial services (Alemu and Ganewo, 2022); and (2) organizational barriers that prevent firms from adopting new technologies (Atkin et al., 2017) and using inputs optimally. In this light, the policy implication of these findings is that, besides easier access to finance, interventions should be aimed at training managers and employees on the use of new technologies and better managerial practices to increase the formal access to credit and public services, which are key to SMEs' growth.

The complexity of interventions to create jobs requires close collaboration among scientists, stakeholders, and practitioners, since the impact is closely related to how research is organized and the intervention is implemented (Taverdet-Popiolek, 2022). In this regard, researchers scarcely provide information about the process evaluation. Process evaluations focus on how a program is implemented and operates, assessing whether it conforms to its original design and documenting barriers on its development and operation. Evidence from process evaluations can complement impact evaluation results and provide a more complete picture of program performance, shedding light on how processes are functioning such as risks and barriers to accomplish the planning. This is particularly important in building the sample: While project beneficiaries can be held accountable to respond to the survey, response rates are likely to be lower among non-beneficiaries. Increasing response rates among non-beneficiaries will involve creating sufficient incentives for them to participate in the baseline and ex-post surveys.

## Future Research

We propose the following research questions to be addressed in future studies more deeply in order to get more robust results to better answer the challenging question of how to create more jobs:

- Do researchers report more transparently the results of the experiments (such as randomization of participants, matching technique, effect size, power analysis, confidence intervals, time effects, budget allocation)?

- Do researchers and practitioners work cooperatively to design the research, monitor the implementation of the intervention, and do the process evaluation?
- Are the effects of matching grants consistent across business sizes, business ages, entrepreneurs ages, gender, sectors, urban vs. rural business location, peaceful and stable vs. fragile and conflict environments, technology usage, and number of business partners and diverse business network?
- Are the samples and the budget allocation clearly defined for each research question and intervention?
- Are the counterfactual and spillover effects sufficiently evaluated?
- Is there any potential nonlinearity of policy impact on job creation?

All these research questions could reduce the inefficiency of policies in order to boost the expected outcomes in job creation.

**Author Contributions** All authors contributed to the study conception and design. Material preparation, data collection, and analysis were performed by Paloma Bernal-Turnes, Ricardo Ernst, and Ana Vico Belmonte. The first draft of the manuscript was written by Paloma Bernal-Turnes, and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

**Funding** Open Access funding provided thanks to the CRUE-CSIC agreement with Springer Nature.

**Data Availability** Not applicable.

**Code Availability** Not applicable.

## Declarations

**Ethics Approval** All authors respect the rules of the journal and the compliance with ethical standards. Authors assume all ethical responsibilities from the authorship and those established by the COPE guidelines.

**Consent to Participate** All authors agreed with the content; all gave explicit consent to submit, and they obtained consent from the responsible authorities at the institute/organization where the work has been carried out before the work is submitted.

**Consent for Publication** Please see the relevant sections in the submission guidelines for further information as well as various examples of wording. Please revise/customize the sample statements according to your own needs.

**Conflict of Interest** The authors declare no competing interests.

**Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

## References

- Aguinis, H., & Bradley, K. J. (2014). Best practice recommendations for designing and implementing experimental vignette methodologies studies. *Organizational Research Methods*, 17(4), 351–371. <https://doi.org/10.1177/1094428114547952>
- Aguinis, H., & Edwards, J. R. (2014). Methodological wishes for the next decade and how to make wishes come true. *Journal of Management Studies*, 51(1), 143–174. <https://doi.org/10.1111/joms.12058>
- Aguinis, H., & Lawal, S. O. (2012). Conducting field experiments using e-Lancing's natural environment. *Journal of Business Venturing*, 27(4), 493–505. <https://doi.org/10.1016/j.jbusvent.2012.01.002>
- Aguinis, H., Pierce, C. A., Bosco, F. A., & Muslin, I. S. (2009). First decade of organizational research methods: Trends in design, measurement, and data-analysis topics. *Organizational Research Methods*, 12(1), 69–112. <https://doi.org/10.1177/1094428108322641>
- Alemu, A., & Ganewo, Z. (2022). Impact analysis of formal microcredit on income of borrowers in rural areas of sidama region, Ethiopia: A propensity score matching approach. *Journal of the Knowledge Economy*. <https://doi.org/10.1007/s13132-021-00863-1>
- Allen, D. G., Hancock, J. I., Vardaman, J. M., & McKee, D. N. (2014). Analytical mindsets in turnover research. *Journal of Organizational Behavior*, 35(S1), S61–S86. <https://doi.org/10.1002/job.1912>
- Ashta, A., Ghosh, C., Guha, S., & Lentz, F. (2021). Knowledge in microsocioal milieu: The case of microfinance practices among women in India. *Journal of the Knowledge Economy*, 12, 146–165. <https://doi.org/10.1007/s13132-016-0372-x>
- Atkin, D., Khandelwal, A. K., & Osman, A. (2017). Exporting and firm performance: Evidence from a randomized experiment. *The Quarterly Journal of Economics*, 132(2), 551–615. <https://doi.org/10.1093/qje/qjx002>
- Atkinson, A. C., & Donev, A. N. (1992). *Optimum experimental designs*. Oxford University Press.
- Austin, J., Scherbaum, C., & Mahlman, R. A. (2002). History of research methods in industrial organizational psychology: Measurement, design, analysis. In M. A. Malden (Ed.), *Handbook of Research Methods in Industrial and Organizational Psychology* (pp. 77–98). Blackwell. <https://doi.org/10.1002/9780470756669.ch1>
- Ayyagari, M., Beck, T., & Demirguc-Kunt, A. (2007). Small and medium enterprises across the globe. *Small Business Economics*, 29, 415–434. <https://doi.org/10.1007/s11187-006-9002-5>
- Ayyagari, M., Demirguc-Kunt, A., & Maksimovic, V. (2011). *Small vs. young firms across the world*. In *Contribution to employment, job creation and growth*. Policy Research Working Paper, 5631. The World Bank.
- Bernal-Turnes, P., & Ernst, R. (2015). Strategies to measure direct and indirect effects in multi-mediator models. *China-USA Business Review*, 14(10), 504–514. <https://doi.org/10.17265/1537-1514/2015.10.003>
- Bernal-Turnes, P., & Ernst, R. (2016). The use of longitudinal mediation models for testing causal effects and measuring direct and indirect effects. *China-USA Business Review*, 15(1), 1–13. <https://doi.org/10.17265/1537-1514/2016.01.001>
- Bezeau, S., & Graves, R. (2001). Statistical power and effect sizes of clinical neuropsychology. *Journal of Clinical and Experimental Neuropsychology*, 23(3), 399–406. <https://doi.org/10.1076/jcen.23.3.399.1181>
- Biscotti, A. M., & D'Amico, E. (2019). Does equity market differently perceive IC management and disclosure behaviours? *Journal of Knowledge Economy*, 10, 756–775. <https://doi.org/10.1007/s13132-017-0492-y>
- Borsboom, D., Kievit, R. A., Cervone, D., & Hood, S. B. (2009). The two disciplines of scientific psychology, or: The disunity of psychology as a working hypothesis. In J. Valsiner, P. Molenaar, M. Lyra, & N. Chaudary (Eds.), *Dynamic process Methodology in the Social and Developmental Sciences* (pp. 67–98). Springer. [https://doi.org/10.1007/978-0-387-95922-1\\_4](https://doi.org/10.1007/978-0-387-95922-1_4)
- Buba, J., Gonzalez, A., & Rizvi, A. (2020). Empirical evidence on firm growth and jobs in developing countries. In *Jobs Working Paper*. The World Bank <https://openknowledge.worldbank.org/handle/10986/34958>.
- Bullock, J. G., Green, D. P., & Ha, H. E. (2010). Yes, but what's the mechanism? (Don't expect an easy answer). *Journal of Personality and Social Psychology*, 98(4), 550. <https://doi.org/10.1037/a0018933>

- Bruhn, M., Karlan, D., & Schoar, A. (2010). What capital is missing in developing countries? *American Economic Review*, *100*(2), 629–633. <https://doi.org/10.1257/aer.100.2.629>
- Bruhn, M., & Love, I. (2012). The real impact of improved access to finance: Evidence from Mexico. *Emerging Markets: Finance eJournal*. <https://doi.org/10.1111/jofi.12091>
- Campbell, D. T. (1969a). Prospective: Artifact and control. In R. Rosenthal & R. L. Rosnow (Eds.), *Artifact in Behavioral Research* (pp. 264–286). Oxford University Press.
- Campbell, D. T. (1969b). Reforms as experiments. *American Psychologist*, *24*(4), 409–429. <https://doi.org/10.1037/h0027982>
- Campbell, D. T., & Stanley, J. L. (1966). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.
- Cascio, W. F., & Zedeck, S. (1983). Open a new window in rational research planning: Adjust alpha to maximize statistical power. *Personnel Psychology*, *36*(3), 517–526. <https://doi.org/10.1111/j.1744-6570.1983.tb02233.x>
- Cohen, J. (1962). The statistical power of abnormal social psychological research: A review. *Journal of Abnormal and Social Psychology*, *65*(3), 145. <https://doi.org/10.1037/h0045186>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Routledge.
- Cohen, J. (1994). The earth is round ( $p < .05$ ). *American Psychologist*, *49*(12), 997. <https://doi.org/10.1037/0003-066X.49.12.997>
- Cohen, J. (1992). Statistical power analysis. *Current Directions in Psychological Science*, *1*(3), 98–101. <https://doi.org/10.1111/1467-8721.ep10768783>
- Cohn, L. D., & Becker, B. J. (2003). How meta-analysis increases statistical power. *Psychological Methods*, *8*(3), 243. <https://doi.org/10.1037/1082-989X.8.3.243>
- Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.
- Cook, T. D., & Steiner, P. M. (2010). Case matching and the reduction of selection bias in quasi-experiments: The relative importance of pretest measures of outcomes, of unreliable measurement, and of mode of data analysis. *Psychological Methods*, *15*(1), 56. <https://doi.org/10.1037/a0018536>
- Cravo, T. A., Gourlay, A., & Becker, B. (2012). SMEs and regional economic growth in Brazil. *Small Business Economics*, *38*, 217–230. <https://doi.org/10.1007/s11187-010-9261-z>
- Cravo, T. A., & Piza, C. (2016). *The impact of business support services for small and medium enterprises on firm performance in low- and middle-income countries. A meta-analysis*. Policy Research Working Paper, 7664. World Bank Accessed January 30, 2023, from <http://hdl.handle.net/10986/24501>.
- Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist*, *30*(2), 116. <https://doi.org/10.1037/h0076829>
- Davis, G. F. (2010). Do theories of organizations progress? *Organizational Research Methods*, *13*(4), 690–709. <https://doi.org/10.1177/1094428110376995>
- Dvouletý, O., Srhoj, S., & Pantea, S. (2021). Public SME grants and firm performance in European Union: A systematic review of empirical evidence. *Small Business Economics*, *57*, 243–263. <https://doi.org/10.1007/s11187-019-00306-x>
- Dubin, R. (1976). Theory building in applied areas. In M. Dunnette (Ed.), *Handbook of industrial and organizational psychology* (pp. 17–39). Rand McNally College.
- Eden, D. (2017). Field experiments in organizations. *Annual Review of Organizational Psychology and Organizational Behavior*, *4*, 91–122. <https://doi.org/10.1146/annurev-orgpsych-041015-062400>
- Eden, D., Stone-Romero, E. F., & Rothstein, H. R. (2015). Synthesizing results of multiple randomized experiments to establish causality in mediation testing. *Human Resource Management Review*, *25*(4), 342–351. <https://doi.org/10.1016/j.hrmr.2015.02.001>
- Edwards, J. R. (2008). To prosper organizational psychology should ... overcome methodological barriers to progress. *Journal of Organizational Behavior*, *29*(4), 469–491. <https://doi.org/10.1002/job.529>
- Faul, F., Erdfelder, E., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. <https://doi.org/10.3758/BRM.41.4.1149>
- Fiala, N. (2018). Returns to microcredit, cash grants and training for male and female microentrepreneurs in Uganda. *World Development*, *105*, 189–200. <https://doi.org/10.1016/j.worlddev.2017.12.027>
- Gelman, A., & Hill, J. (2006). *Data analysis using regression and multilevel/hierarchical models. (Analytical Methods for Social Research)*. Cambridge University Press. <https://doi.org/10.1017/CBO9780511790942>

- Gergen, K. J. (1973). Social psychology as history. *Journal of Personality and Social Psychology.*, 26(2), 309. <https://doi.org/10.1037/h0034436>
- Gergen, K. J. (1982). Toward transformation in social knowledge. In F. Kidd (Ed.), *Springer Series in Social Psychology*. Springer-Verlag. <https://doi.org/10.1007/978-1-4612-5706-6>
- Gertler, P. J., Martinez, S., Premand, P., Rawlings, L. B., & Vermeersch, C. M. J. (2016). *Impact evaluation in practice*. Inter-American Development Bank and World Bank. <https://doi.org/10.1596/978-1-4648-0779-4>
- Green, P., & MacLeod, C. J. (2016). SIMR: An R package for power analysis of generalized linear mixed models by simulation. *Methods in Ecology and Evolution*, 7(4), 493–498. <https://doi.org/10.1111/2041-210X.12504>
- Gibbons, R., & Henderson, R. (2012). Relational contracts and organizational capabilities. *Organization Science*, 23(5), 1350–1364. <https://doi.org/10.1287/orsc.1110.0715>
- Grant, A. M., & Wall, T. D. (2009). The neglected science and art of quasi-experimentation. Why-to, when-to, how-to advice to organizational researchers. *Organizational Research Methods.*, 12(4), 653–686. <https://doi.org/10.1177/1094428108320737>
- Gregorie, D. A., Shepherd, D. A., & Lambert, L. S. (2010). Measuring opportunity recognition beliefs. Illustrating and validating an experimental approach. *Organizational Research Methods*, 13(1), 114–145. <https://doi.org/10.1177/1094428109334369>
- Grimm, M., & Paffhausen, A. (2015). Do interventions targeted at micro-entrepreneurs and small and medium-sized firms create jobs? A systematic review of the evidence for low and middle income countries. *Labour Economics.*, 32, 67–85. <https://doi.org/10.1016/j.labeco.2015.01.003>
- Hall, B., & Khan, B. (2003). *Adoption of new technology*. National Bureau of Economic Research. NBER Working Paper Accessed December 16, 2022, from <http://www.nber.org/papers/w9730>
- Halpern, S. D., Karlawish, J. H. T., & Berlin, J. A. (2002). The continuing unethical conduct of underpowered clinical trials. *The Journal of the American Medical Association.*, 288(3), 358–362. <https://doi.org/10.1001/jama.288.3.358>
- Haltiwanger, J., Jarmin, R. S., & Miranda, J. (2013). Who creates jobs? Small versus large versus young. *Review of Economics and Statistics.*, 95(2), 347–361. [https://doi.org/10.1162/REST\\_a\\_00288](https://doi.org/10.1162/REST_a_00288)
- Hedges, L. V. (1987). How hard is hard science, how soft is soft science? The empirical cumulativeness of research. *American Psychologist*, 42(5), 443. <https://doi.org/10.1037/0003-066X.42.5.443>
- Hedges, L. V., & Olkin, I. (1980). Vote-counting methods in research synthesis. *Psychological Bulletin.*, 88(2), 359. <https://doi.org/10.1037/0033-2909.88.2.359>
- Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Academic Press.
- Histrova, D. E., & Coste, A. (2016). *How to make grants a better match for private sector*. World Bank <https://openknowledge.worldbank.org/handle/10986/26434>
- Hitt, M. A., Boyd, B. K., & Li, D. (2004). The state of strategic management research and a vision of the future. *Research Methodology in Strategy and Management*. [https://doi.org/10.1016/S1479-8387\(04\)01101-4](https://doi.org/10.1016/S1479-8387(04)01101-4)
- Highhouse, S. (2009). Designing experiments that generalize. *Organizational Research Methods.*, 12(3), 554–566. <https://doi.org/10.1177/1094428107300396>
- Holm, H. J., Opper, S., & Nee, V. (2013). Entrepreneurs under uncertainty: An economic experiment in China. *Management Science.*, 59(7), 1671–1687. <https://doi.org/10.1287/mnsc.1120.1670>
- Howard, G. S., Maxwell, S. E., & Fleming, K. J. (2000). The proof of the pudding: An illustration of the relative strengths of null hypothesis, meta-analysis, and Bayesian analysis. *Psychological Methods*. <https://doi.org/10.1037/1082-989x.5.3.315>
- Hsu, D. K., Haynie, J. M., Simmons, S. A., & McKelvie, A. (2014). What matters, matters differently: A conjoint analysis of the decision policies of angel and venture capital investors. *Venture Capital.*, 16(1), 1–25. <https://doi.org/10.1080/13691066.2013.825527>
- Hsu, D. K., Simmons, S. A., & Wieland, A. M. (2017). Designing entrepreneurship experiments: A review, typology, and research agenda. *Organizational Research Methods.*, 20(3), 379–412. <https://doi.org/10.1177/1094428116685613>
- Hunter, J. E., & Schmidt, F. L. (1990). *Methods of meta-analysis: Correcting error and bias in research findings*. Thousand Oaks.
- Johns, G. (2006). The essential impact of context on organizational behavior. *Academy of Management Review.*, 31(2), 386–408. <https://doi.org/10.2307/20159208>
- Judd, C. M., & McClelland, G. H. (1989). *A model comparison approach to regression, ANOVA, and beyond*. Routledge. <https://doi.org/10.4324/9781315744131>

- Kaiser, H. F. (1960). Directional statistical decisions. *Psychological Review*, 67(3), 160. <https://doi.org/10.1037/h0047595>
- Karlan, D., Knight, R., & Udry, C. (2015). Consulting and capital experiments with microenterprise tailors in Ghana. *Journal of Economic Behavior and Organization*, 118, 281–302. <https://doi.org/10.1016/j.jebo.2015.04.005>
- Ketchen, D. J., Boyd, B. K., & Bergh, D. D. (2008). Research methodology in strategic management: Past accomplishments and future challenges. *Organizational Research Methods*, 11(4), 643–658. <https://doi.org/10.1177/1094428108319843>
- Khandker, S. R., Koolwal, G. B., & Samad, H. A. (2010). *Handbook on impact evaluation: Quantitative methods and practices*. The World Bank Accessed January 30, 2023, from <https://openknowledge.worldbank.org/handle/10986/2693>.
- King, E. B., Hebl, M. R., Morgan, W. B., & Ahmad, A. S. (2012). Field experiments on sensitive organizational topics. *Organizational Research Methods*. <https://doi.org/10.1177/1094428112462608>
- Knight, F. H. (1921). *Risk, uncertainty and profit*. Houghton Mifflin.
- Koudstaal, M., Sloof, R., & van Praag, C. M. (2015). Risk, uncertainty, and entrepreneurship: Evidence from a large lab-in-the-field experiment. *Management Science*. <https://doi.org/10.1287/mnsc.2015.2249>
- Kozlowski, S. W. J., & Klein, K. J. (2000). A multilevel approach to theory and research in organizations: Contextual, temporal, and emergent processes. In K. J. Klein & S. W. J. Kozlowski (Eds.), *Multilevel theory, research, and methods in organizations: Foundations, extensions, and new directions* (pp. 3–90). Jossey-Bass.
- Kruskal, W. (1978). Taking data seriously. In Y. Elkana, J. Lederberg, R. K. Merton, A. Thackray, & H. Zuckerman (Eds.), *Toward a metric of science: The advent of science indicators* (pp. 139–171). John Wiley & Sons.
- Leventhal, L., & Huynh, C. L. (1996). Directional decisions for two-tailed tests: Power, error rates, and sample size. *Psychological Methods*, 1(3), 278. <https://doi.org/10.1037/1082-989X.1.3.278>
- Legovini, A., Maro, D., & V., Piza, C. (2015). *Impact evaluation helps deliver development projects*. Policy Research Working Paper. The World Bank. <https://doi.org/10.1596/1813-9450-7157>
- Lopez-Acevedo, G., & Tan, H. W. (2011). *Impact evaluation of small and medium enterprise programs in Latin America and Caribbean*. The World Bank. <https://doi.org/10.1596/978-0-8213-8775-7>
- Lopez-Acevedo, G., & Tinajero, M. (2010). *Mexico: Impact evaluation of SME programs using panel firm data*. Policy Research Working Paper Series. The World Bank. <https://doi.org/10.1596/1813-9450-5186>
- Lykken, D. T. (1968). Statistical significance in psychological research. *Psychological Bulletin*, 70(3p1), 151. <https://doi.org/10.1037/h0026141>
- Maddock, J. E., & Rossi, J. S. (2001). Statistical power of articles published in three health psychology-related journals. *Health Psychology*, 20(1), 76. <https://doi.org/10.1037/0278-6133.20.1.76>
- Mallett, R., & Slater, R. (2016). Livelihoods, conflict and aid programming: Is the evidence base good enough? *Disasters*, 40(2), 226–245. <https://doi.org/10.1111/disa.12142>
- Mano, Y., Iddrisu, A., Yoshino, Y., & Sonobe, T. (2012). How can micro and small enterprises in Sub-Saharan Africa become more productive? The impacts of experimental basic managerial training. *World Development*, 40(3), 458–468. <https://doi.org/10.1016/j.worlddev.2011.09.013>
- Maxwell, S. E. (2004). The persistence of underpowered studies in psychological research: Causes, consequences, and remedies. *Psychological Methods*, 9(2), 147. <https://doi.org/10.1037/1082-989X.9.2.147>
- Maxwell, S. E., Delaney, H. D., & Kelley, K. (2018). *Designing experiments and analyzing data: A model comparison perspective*. Routledge.
- McCallum, R. C., Browne, M. W., & Sugawara, H. M. (1996). Power analysis and determination of sample size for covariance structure modeling. *Psychological Methods*, 1(2), 130. <https://doi.org/10.1037/1082-989X.1.2.130>
- McClelland, G. H. (1997). Optimal design in psychological research. *Psychological Methods*, 2(1), 3. <https://doi.org/10.1037/1082-989X.2.1.3>
- McKelvie, A., Haynie, J. M., & Gustavsson, V. (2011). Unpacking the uncertainty construct: Implications for entrepreneurial action. *Journal of Business Venturing*, 26(3), 273–292. <https://doi.org/10.1016/j.jbusvent.2009.10.004>

- McKenzie, D., Assaf, N., & Cusolito, A. P. (2015). The additionality impact of a matching grant program for small firms: Experimental evidence from Yemen. *Policy Research Working Paper*. <https://doi.org/10.1596/1813-9450-7462>
- McKenzie, D., Assaf, N., & Cusolito, A. P. (2016). The demand for, and impact of, youth internships: Evidence from a randomized experiment in Yemen. *IZA Journal of Labor*, 5(1), 1–15. <https://doi.org/10.1186/s40175-016-0048-8>
- McMullen, J. S., & Sheperd, D. A. (2006). Entrepreneurial action and the role of uncertainty in the theory of entrepreneur. *Academy of Management Review*, 31(1), 132–152. <https://doi.org/10.5465/AMR.2006.19379628>
- Mead, L. M. (1988). Welfare policy: The administrative frontier. *Journal of Policy Analysis and Management*. [https://doi.org/10.1002/\(SICI\)1520-6688\(199623\)15:4<587::AID-PAM5>3.0.CO;2-D](https://doi.org/10.1002/(SICI)1520-6688(199623)15:4<587::AID-PAM5>3.0.CO;2-D)
- Mead, D. C., & Leidholm, C. (1998). The dynamics of micro and small enterprises in developing countries. *World Development*, 26(1), 61–74. [https://doi.org/10.1016/S0305-750X\(97\)10010-9](https://doi.org/10.1016/S0305-750X(97)10010-9)
- Meehl, P. E. (1978). Theoretical risks and tabular asterisks: Sir Karl, Sir Ronald and the slow progress of soft psychology. *Journal of Consulting and Clinical Psychology*. <https://doi.org/10.1037/0022-006X.46.4.806>
- Miller, K. D., & Tsang, E. W. (2011). Testing management theories: Critical realist philosophy and research methods. *Strategic Management Journal*, 32(2), 139–158. <https://doi.org/10.1002/smj.868>
- Martínez, M. (2022). Competitive advantage and knowledge absorptive capacity: The mediating role of innovative capability. *Journal of the Knowledge Economy*, 13(1), 185–210. <https://doi.org/10.1007/s13132-020-00708-3>
- Mone, M. A., Mueller, G. C., & Mauland, W. (1996). The perceptions and usage of statistical power in applied psychology and management research. *Personnel Psychology*, 49(1), 103–120. <https://doi.org/10.1111/j.1744-6570.1996.tb01793.x>
- Pfeffer, J. (1993). Barriers to the advance of organizational science: Paradigm development as a dependent variable. *Academy of Management Review*, 18(4), 599–620. <https://doi.org/10.5465/amr.1993.9402210152>
- Piza, C., Cravo, T., Taylor, L., Gonzalez, L., Musse, I., Furtado, I., Sierra A. C., & Abdelnour, S. (2016). The impact of business support services for small and medium enterprises on firm performance in low and middle-income countries: A systematic review. *Campbell Systematic Reviews*. 12(1), 1–167. <https://doi.org/10.4073/csr.2016.1>
- Podsakoff, P. M., & Dalton, D. R. (1987). Research methodology in organizational studies. *Journal of Management*, 13(2), 419–441. <https://doi.org/10.1177/014920638701300213>
- Podsakoff, N. P., Podsakoff, P. M., MacKenzie, S. B., Maynes, T. D., & Spoelma, T. M. (2014). Consequences of unit-level organizational citizenship behaviors: A review and recommendations for future research. *Journal of Organizational Behavior*, 35(S1), S87–S119. <https://doi.org/10.1002/job.1911>
- Ralston, L. (2014). *Job creation in fragile and conflict-affected situations*. Policy Research Working Paper. <https://doi.org/10.1596/1813-9450-7078>
- Rijkers, B., Ruggeri, C., & Teal, F. (2010). Who benefits from promoting small enterprises? Some empirical evidence from Ethiopia. *World Development*, 38(4), 523–540. <https://doi.org/10.1016/j.worlddev.2009.10.00>
- Rosenbaum, P. R. (2002). *Observational studies*. Springer.
- Rosenthal, R. (1991). *Meta-analytic procedures for social research*. SAGE Publications. <https://doi.org/10.4135/9781412984997>
- Rossignol, I., & Salmon, K. (2016). *Stimulating the private sector and job creation in fragile and conflict-affected settings*. Position Paper Trade and Competitiveness Global Practice Accessed January 30, 2023, from <https://openknowledge.worldbank.org/bitstream/handle/10986/25296/108777-WP-P156896-PUBLIC-ABSTRACT-SENT-IntegratedFrameworkforJobsinFragileandConflictSituation.sfinal.txt?sequence=2&isAllowed=y>
- Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5), 688. <https://doi.org/10.1037/h0037350>
- Rubin, D. B. (2007). The design versus the analysis of observational studies for causal effects: Parallels with the design of randomized trials. *Statistics in Medicine*, 26(1), 20–36. <https://doi.org/10.1002/sim.2739>
- Rubin, D. B. (2008). For objective causal inference, design trumps analysis. *The Annals of Applied Statistics*, 2(3), 808–840. <https://doi.org/10.1214/08-AOAS187>

- Rynes, S. L., & Bartunek, J. M. (2017). Evidence-based management: Foundations, development, controversies, and future. *Annual Review of Organizational Psychology and Organizational Behavior*, 4, 235–261. <https://doi.org/10.1146/annurev-orgpsych-032516-113306>
- Satorra, A., & Saris, W. E. (1985). The power of the likelihood ratio test in covariance structure analysis. *Psychometrika*, 50, 83–90. <https://doi.org/10.1007/BF02294150>
- Saris, W. E., & Satorra, A. (1993). Power evaluations in structural equation models. In K. A. Bollen & J. S. Long (Eds.), *Testing Structural Equation Models* (pp. 181–204). SAGE Publications.
- Sawyer, A., & Ball, A. (1981). Statistical power and effect size in marketing research. *Journal of Marketing Research*, 18(3), 275–290. <https://doi.org/10.1177/002224378101800302>
- Scandura, T. A., & Williams, E. A. (2000). Research methodology in management: Current practices, trends, and implications for future research. *Academy of Management Journal*, 43(6), 1248–1264. <https://doi.org/10.5465/1556348>
- Schmidt, F. L., Hunter, J. E., & Urry, V. W. (1976). Statistical power in criterion-related validity studies. *Journal of Applied Psychology*, 61(4), 473–485. <https://doi.org/10.1037/0021-9010.61.4.473>
- Schmitt, N. W., Klimoski, R. J., Ferris, G. R., & Rowland, K. M. (1991). *Research methods in human resource management*. South-Western Publishing.
- Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). *Experimental and quasi-experimental designs for generalized causal inference*. Houghton Mifflin.
- Shane, S. (2003). *A general theory of entrepreneurship: The individual–opportunity nexus*. Edward Elgar Publishing.
- Shane, S. A., & Venkataraman, S. (2000). The promise of entrepreneurship as a field of research. *Academy of Management Review*, 25(1), 217–226. <https://doi.org/10.5465/amr.2000.2791611>
- Smith, F. J. (1977). Work attitudes as predictors of attendance on a specific day. *Journal of Applied Psychology*, 62(1), 16. <https://doi.org/10.1037/0021-9010.62.1.16>
- Spector, P. E. (1981). *Research designs*. SAGE Publications. <https://doi.org/10.4135/9781412985673>
- Sutcliffe, J. P. (1980). On the relationship of reliability to statistical power. *Psychological Bulletin*, 88(2), 509. <https://doi.org/10.1037/0033-2909.88.2.509>
- SPSS. (2017). *Sample Power Manual*. SPSS.
- Tan, H., & Lopez-Acevedo, G. (2005). Evaluating training programs for small and medium enterprises: Lessons from Mexico. Policy Research Working Paper. World Bank. <https://doi.org/10.1596/1813-9450-3760>
- Taverdet-Popiolek, N. (2022). Economic footprint of a large french research and technology organisation in Europe: Deciphering a simplified model and appraising the results. *Journal of the Knowledge Economy*, 13(1), 44–69. <https://doi.org/10.1007/s13132-020-00709-2>
- Tversky, A., & Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin*, 76(2), 105–110. <https://doi.org/10.1037/h0031322>
- Uy, M. A., Foo, M. D., & Aguinis, H. (2010). Using experience sampling methodology to advance entrepreneurship theory and research. *Organizational Research Methods*, 13(1), 31–54. <https://doi.org/10.1177/1094428109334977>
- West, S. G., & Thoemmes, F. (2010). Campbell's and Rubin's perspectives on causal inference. *Psychological Methods*, 15(1), 18. <https://doi.org/10.1037/a0015917>
- Whetten, D. A. (1989). What constitutes a theoretical contribution? *Academy of Management Review*, 14(4), 490–495. <https://doi.org/10.2307/258554>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.