the standard methodological tool set meant a departure from the other class. The social sciences are still generating numerous and important research findings using quasiexperiments, while increasingly also carrying out experiments; the health sciences are unlikely to decrease the use of experiments in establishing treatment efficacy of novel technologies as the increasingly realize quasi-experimental opportunities for causal inference. Rather, the adoption of a new methods class opens up opportunities to answer research questions that could not be answered with the previously used class of methods. For instance, in the health sciences, an important driver of the use of quasi-experimental methods has been the "impact evaluation" agenda, which focuses on how to establish the "real-life" impact of medical technologies whose efficacy has been established in randomized controlled experiments (Bärnighausen, et al., 2017).

In particular, the potential quasi-experiments have to provide strong causal evidence when RCTs are not possible; the strength of quasi-experiments in producing externally valid results; and the potential quasi-experiments have to establish causal effects on long-term and non-health outcomes should be considered when debating whether to include quasi-experimental results in the synthesis of a body of evidence. Quasiexperiments are typically graded as observational, non-experimental studies, but this assessment does not allow that some quasi-experiments, e.g. regression discontinuity designs and randomized natural or policy experiments, yield inferences nearly as strong as randomized clinical trials. Furthermore, the features of a rigorously-presented quasiexperimental study differ from the features of a rigorously-presented cohort or case-control study.

Devising checklists for the reporting and assessment of quasi-experiments is difficult due to the wide range of data generating processes underlying quasi-experiments and the wide range of methods used to analyze them. Often, specific subject area knowledge is required to determine the plausibility of assumptions invoked in a quasi-experiment. And yet, the potential of these designs to produce rigorous, actionable evidence to improve patient care, policy design, and resource allocation is too great to be ignored. One important area for future research is how to best integrate quasi-experiments into methods for evidence synthesis. Stronger integration of quasi-experimental results into evidence synthesis may have powerful behavioral feedback effects. Quasi-experiments have great potential to generate novel and important insights, but it is likely that they are currently substantially underutilized relative to this potential. Increasing use of quasi-experimental results in evidence synthesis may contribute to closing the gap between the potential and the realized

contributions of quasiexperimental studies in informing health care practice, programs and policy (Goldfarb & Tucker, 2014).

2. Types of quasi- experimental approaches

There are many different techniques for creating a valid comparison group. Commonly mentioned in development literature, and these are Non-equivalent Groups Design (NEGD), Propensity Score Matching (PSM), Regression Discontinuity Design (RDD), Reflexive Comparisons and Time series design.

2.1 Non-equivalent Groups Design (NEGD)

NEGD is the most frequently used quasi-experimental approach used in the social sciences, and is certainly the most common method used by CSOs. The aim is to identify comparison groups that are as similar as possible to the target population. But the comparison groups normally exist as groups prior to the development intervention. For example, if children in a school or classroom form part of a target population than a comparison group could be developed from a similar school or classroom. If working with communities then a comparison group could be developed from people living in a village that is not targeted by a project or programme but is very similar to one that is (White & Sabarwal, 2014) .

The main drawback of NEGD is that it is never possible to be sure that the intervention and comparison groups are entirely similar, which is why studies based on NEGD are often less reliable and require more careful interpretation than studies based on RCTs. In other words, prior differences between the groups could affect differences in results measured at a later date. Studies based on NEGD almost always consist of a baseline and follow-up, so differences at baseline can be assessed as well as differences during or after an intervention (Intrac, 2017).

2.2 Propensity Score Matching (PSM)

PSM is a common method used to select a comparison group after data collection has taken place. It attempts to directly match individual units (individuals, households, organisations etc.) that have received an intervention, with those that have not. Ideally, it would be possible to directly match units according to different characteristics. For example, if a household in an intervention group consisted of a husband and wife, aged between 30-40, with two boys and a girl, and owning two hectares of land and three cows, then it would be ideal to have an exactly similar household in the comparison group (Banerjee & Duflo, 2009).

In PSM, an individual is not matched on every single observable characteristic, but on their propensity score – that is, the likelihood that the individual will participate in the intervention (predicted likelihood of participation) given their observable characteristics. PSM thus matches treatment individuals/households with similar comparison individuals/households, and subsequently calculates the average difference in the indicators of interest. In other words, PSM ensures that the average characteristics of the treatment and comparison groups are similar, and this is deemed sufficient to obtain an unbiased impact estimate (White & Sabarwal, 2014).

In practice this is not usually possible to do. Instead, PSM uses a set of statistical analysis techniques to create a comparison group that is as similar as possible to the sample in the target population, across all the different characteristics. This results in the formation of two groups that have similar average characteristics. The assumption is that the groups are therefore close enough that results will not be biased (Intrac, 2017).

PSM involves the following five steps:

1. **Ensure representativeness** – Ensure that there is a representative sample survey of eligible participants and non-participants in the intervention. Baseline data are preferred for calculating propensity scores. This technique can, however, also be used with endline data: the matching variables must be variables that are unaffected by the intervention.
2. **Estimate propensity scores** – The propensity scores are constructed using the 'participation equation', which is either a log it or probit regression with programme participation as the dependent variable (in the programme = 1, not in the programme = 0). The characteristics deemed to affect participation should be well considered and as exhaustive as possible, but should exclude characteristics that may have been affected by the intervention. For this reason, it is best to use baseline data, where available, to estimate the propensity scores.

3. **Select a matching algorithm** – Each member of the treatment group is then matched to one or more members of the comparison group. There are different ways of doing this such as matching each participant to their 'nearest neighbor' non-participant. The mean of the closest five neighbors is most commonly used. A single individual in the comparison group may be matched to several different individuals in the treatment group. In order for the matching to be valid, it is essential to compare 'observed values' for

participants and non-participants with the same range of characteristics. Observations in the comparison group with a propensity score lower than the lowest observed value in the treatment group are discarded. Similarly, observations in the treatment group with a propensity score higher than the highest observed value in the comparison group are also discarded.

4. **Check for balance** – The characteristics of the treatment and comparison groups are compared to test for balance. Ideally, there will be no significant differences in average observable characteristics between the two groups. Now that the treatment and comparison groups are similar on observable characteristics, variance in the incidence of child diarrhea between the treatment and comparison groups can be attributed to differences such as access to clean water.

5. **Estimate programme effects and interpret results** – Finally, the impact estimate, either single or double difference, is calculated by firstly calculating the difference between the indicator for the treatment individual and the average value for the matched comparison individuals, and secondly averaging out all of these differences (White & Sabarwal, 2014).

2.2.1 Advantages of PSM

The two main advantages of PSM are that it is always feasible if data are available, and it can be done after an intervention has finished, including in the absence of baseline data (although this is not ideal). If baseline data are unavailable, 'recall' can be used to reconstruct pre-intervention characteristics. This can be imprecise, however, and common sense should prevail when deciding which variables can be recalled accurately.

2.2.2. Disadvantage of PSM

The main drawback is that PSM relies on matching individuals on the basis of observable characteristics linked to predicted likelihood of participation. So, if there are any 'unobserved' characteristics that affect participation and which change over time, the estimates will be biased and thus affect the observed results. An additional practical limitation of using PSM is the need for the assistance of a statistician or someone with skills in using different statistical packages.

2.3 Regression Discontinuity Design (RDD)

Regression discontinuity (RD) is another useful quasi-experimental technique in which the 'experiment' relies on an exogenous arbitrary threshold. As Imbens and Lemieux (2008) put

it, the basic idea behind the RD design is that assignment to the treatment is determined, either completely or partly, by the value of the predictor being on either side of a fixed threshold." Hartmann et al. (2011) emphasize the promise of regression discontinuity as an identification strategy for marketing scholarship. They argue that many marketing interventions are based on thresholds of real or expected consumer behavior. For example, direct mail companies use the scoring policies for Recency, Frequency, Monetary (RFM) models. Consumers just above and just below the cutoff should be similar in many dimensions and their outcomes can be compared to assess the impact of the different mailings. Another example is government policies based on firm size. For example, many government policies regarding requirements for firms to post calories, undertake layoffs, and provide benefits depend on the number of employees or other measures of firm size. By comparing firms just above and just below the threshold, it is possible to assess the impact of the policies on firm behavior (Balvin, Hague, & Jackson, 2014).

RDD can only be used when the target population is selected based on meeting a certain threshold (for example, if people only qualify for a project if they are living on less than $1 a day, or have a body-mass index (BMI) of less than 16). In this case, those above and below the threshold may be very different. So, for instance, if looking at prevalence of diseases, a set of people with a BMI of less than 16 could not reasonably be compared with a comparison group of people with a much higher BMI (White & Sabarwal, 2014). This approach can be used when there is some kind of criterion that must be met before people can participate in the intervention being evaluated. This is known as a threshold. A threshold rule determines eligibility for participation in the programme/policy and is usually based on a continuous variable assessed for all potentially eligible individuals. For example, students below a certain test score are enrolled in a remedial programme, or women above or below a certain age are eligible for participation in a health programme (e.g., women over 50 years old are eligible for free breast cancer screening) (Bor, 2014.).

2.3.1 Advantages of RDD

RDD deals with non-observable characteristics more convincingly than other quasi-experimental matching methods. It can also utilize administrative data to a large extent, thus reducing the need for data collection – although the outcome data for those not accepted into the programme often need to be collected (Anderson, 2017).

2.3.2 Disadvantages of RDD

The limits of the technique are that the selection criteria and/or threshold are not always clear and the sample may be insufficiently large for the analysis (as noted above). In addition, RDD yields a 'local area treatment effect'. That is, the impact estimate is valid for those close to the threshold, but the impact on those further from the threshold may be different (it could be more or less). In practice, however, where it has been possible to compare this 'local' effect with the 'average' effect, the differences have not been great (Goldfarb & Tucker, 2014).

2.4 Reflexive comparisons

In a reflexive comparison study, there is no comparison group. A pre- and post- test (baseline and repeat study) is done on a set of units, and the change between the two is attributed to the project intervention. The rationale for calling this a quasi- experimental study is that the units act as their own comparisons. For example, in a project looking to improve farmers' crop yields a sample of the farmers at baseline will not have received any inputs, and can therefore be a comparison group for the same sample of farmers afterwards (Banerjee & Duflo, 2009) . Many CSOs use baselines and follow-up studies to assess change. However, many would be surprised to know that some consider these to be quasi-experimental designs. The main criticism of reflexive comparisons is that they are often unable to distinguish between changes brought about by an intervention and changes due to other effects (Balvin, Hague, & Jackson, 2014).

2.5 Time Series Design

If the group is repeatedly measured before and after the intervention, rather than once before and once after, a time-series design is created. Time-series designs are especially useful when there are continuous, naturally occurring observations of the dependent variable over time and there is a sudden or distinct intervention during the observations. Several conditions should be met in employing this design. These are: The observations should be made at equal time intervals and conducted with the same procedures in order to reduce the threat of instrumentation, The intervention introduced should be a distinctive, abrupt intervention that is clearly new to the existing environment, There should be some evidence that the subjects involved in each observation are the same, There should not be any kind of change affecting the subjects occurring at about the same time as the intervention (Meyer, 1995).

3. Quasi-Experimental Methods for Data Analysis

Once the researcher has found a setting that may help identify the causal effect of interest, the next steps involve exploring the raw data to see the degree to which the quasi-experiment is credible. In particular, the researcher should see that the treatment and control groups are similar in dimensions other than whether they received the treatment. Perhaps the ideal thought experiment here is Zhang (2010), whose treatment and control were a pair of kidneys from the same person. Most research settings are less favorable. Still, researchers should show a comparison of mean values of demographic characteristics and behaviors for the two groups. This comparison should support the argument that these groups are similar. In cases where the treatment occurs in the middle of a time series, demonstrating that the treatment and control groups were similar prior to the arrival of the treatment can be a particularly powerful argument for exogeneity of the treatment. Many papers use a graph that shows that before the treatment occurred, the treatment and control groups were on a similar trend and had similar values; then, after the treatment occurred, the trajectory of the treatment group changed but not the control group (Bor, 2014.).

After establishing similarity between the treatment and control groups in the raw data, the next step is typically to conduct regression analysis that demonstrates the effect of interest. Next, we discuss three different broad identification strategies using quasi-experiments and the process that authors should undertake to convince themselves and their readers that they have identified a causal effect (Goldfarb & Tucker, 2014).

3.1 Single difference impact estimates

Single difference impact estimates compare the outcomes in the treatment group with the outcomes in the comparison group at a single point in time following the intervention

3.2 Difference-in-differences

Difference-in-differences (DID), also known as the ‘double difference’ method, compares the changes in outcome over time between treatment and comparison groups to estimate impact. DID gives a stronger impact estimate than single difference, which only compares the difference in outcomes between treatment and comparison groups following the intervention (at t+1). Applying the DID method removes the difference in the outcome between treatment and comparison groups at the baseline. Nonetheless, this method is best used in conjunction with other matching methods such as PSM or RDD. If DID is used without matching, the researchers should test the ‘parallel trends assumption’, i.e., that the trend in outcomes in

treatment and comparison areas was similar before the intervention. The double difference estimate is greater than the single difference estimate since the comparison group had better WAZ than the treatment group at the baseline. DID allows the initial difference in WAZ between treatment and comparison groups to be removed; single difference does not do this, and so in this example resulted in an underestimate of programme impact.

A key issue in difference-in-differences analysis is correlated errors in observations because the outcome is often observed at a finer level than the treatment. For example, the researcher might observe treatment and control groups for a number of advertising campaigns over a long time period. For each campaign, the researcher might have data on many individuals per campaign and many time periods per individual. It is important to recognize that the choices of the same individual in many time periods are likely to be correlated. Bertrand et al. (2004) emphasized that failure to control for the correlation between these choices will lead to an overstatement of the effective degrees of freedom in the data and therefore standard errors will be biased downwards. They suggest clustering standard errors by individual over time to address this issue and provide Monte Carlo evidence that clustering is likely to lead to robust inference. Similarly, Donald and Lang (2007) emphasize that if individual responses to the same treatment are likely to be correlated (for example, because of close physical or social proximity), clustering standard errors by groups of individuals is a conservative way to estimate standard errors. Researchers often need to decide on the size of the clusters. For example, in studying ready-to-eat breakfast cereals, is the correct unit the company (e.g. General Mills), the brand (e.g. Cheerios), or the subbrand (e.g. Honey Nut Cheerios)? The answer depends on the data and research question. Each cluster should contain those observations most likely to be correlated with each other (Goldfarb & Tucker, 2014).

3.2.1 Advantages and disadvantages of the DID method

The major limitation of the DID method is that it is based on the assumption that the indicators of interest follow the same trajectory over time in treatment and comparison groups. This assumption is known as the ‘parallel trends assumption’. Where this assumption is correct, a programme impact estimate made using this method would be unbiased. If there are differences between the groups that change over time, however, then this method will not help to eliminate these differences (White & Sabarwal, 2014).

3. Presentation of Results and Analysis

When writing up results based on a quasi-experimental evaluation, it is important to provide details about the specific methodology, including data collection Since the success of these methods depends greatly on the quality of data collected (or already available), some sort of assurance of quality should be provided. It is also important to provide information about the tenability of the assumptions on which these methods are based. Although some of the assumptions cannot be tested directly (e.g., parallel trends assumptions) authors should provide clear reasoning as to why they believe these assumptions hold. It is recommended that the description of the methodology includes details of the sampling method as well as the approach to the construction of treatment and comparison groups (including the number of individuals, households or clusters involved). The results can be analysed and reported for the entire sample as well as for important (predefined) subgroups (e.g., by age or by sex) to identify and discuss any differential effects. The findings then need to be linked to the theory of change and used to answer the key evaluation questions (KEQs) – for example, do the findings support the theory of change? If not, which assumption behind the theory of change was not fulfilled? These types of analyses can help evaluators to identify concrete programme or policy recommendations, which should make up the conclusion of the report. In most cases, it would also be useful to include a discussion around whether and to what extent the results can be extrapolated to different settings. Conclusions drawn from quasi-experimental designs are causally valid as long as the assumptions regarding the particular matching method are met. The quality of the match should also be tested and reported (Goldfarb & Tucker, 2014).

5. Strengths and weaknesses of Quasi-Experimental Design

5.1 Strength of Quasi-Experimental Design

Quasi-experimental approaches can provide evidence of change that is more robust than evidence produced without a control or comparison group. It allow CSOs to develop a counterfactual – an estimate of what the situation would have been without the intervention. It can be planned and applied after an intervention has started – unlike RCTs – and can be used in situations where full experimental designs cannot. They are often easier to set up than RCTs, and may require less expertise and resources (Intrac, 2017).

One of the greatest strengths of this approach is its transparency, which contributes to its generalizability and external validity. It is important that researchers ensure they thoroughly scope their research, contextualize the research area and use existing theoretical knowledge to inform their experimental design. This can help them integrate local and collective experience with the scientific rigour of quasi-experiments. Transparency and external validity

help ensure any findings are communicable to both scientific and policy communities, particularly other research institutes, donors, academics and development institutions. Because we tend to use quasi-experimental methods when we need to find answers for practical questions for example, to establish policy or intervention impact these evaluations are often action oriented. Although it is normally outside the mandate of this approach to inform action-oriented change for community-level actors, we can use real-time reflection on interim results to improve intervention design. So findings can inform changes to programming, even if empowerment does not flow down to local communities (Anderson, 2017).

5.2 Weakness of Quasi-Experimental Design

Firstly, in common with RCTs, quasi-experimental approaches attribute changes directly to interventions without considering how the change was produced. Therefore they are unable to always provide explanations of how change came about. This challenge can be partly resolved if alternative, more explanatory methods are used alongside the quasi-experimental approach (Stern, et al., 2012).

Quasi-experimental approaches can help answer the question of what changed over a specific time and in a particular environment. But because they do not investigate how or why changed happened, they cannot always be used to make wider generalizations. Again, this challenge can be resolved by using additional methods where appropriate.

In common with RCTs, quasi-experimental approaches tend to suit interventions where there is a clear, logical link between cause and effect, and where effects are designed to be achieved over short- to medium-term time spans. Yet many CSOs carry out work in highly complex environments, where contribution rather than attribution is considered key, and where links between cause and effect are not always linear.

Quasi-experimental approaches can be difficult and costly to apply, and are often more complex to analyse and interpret than RCTs. This means specialist expertise may be needed. Quasi-experimental approaches tend to be based around the collection of data from large numbers of individuals and households, sometimes over long-time periods. Smaller CSOs may need to buy-in specialist knowledge to design and run them, or may not have the resources to implement them at all (Intrac, 2017).

It is impossible to prove the validity of a quasi-experiment, such as a legitimate control group for another or whether the exclusion restriction holds in instrumental variables. The credibility of any quasi-experimental work therefore relies on the plausibility of the argument for causality rather than on any formal statistical test. Second, external validity depends on the match between the treatments driven by the quasi-experimental variation and the overall sample needed to answer the research question. Quasi-experiments often require a focus on a narrow slice of the data and therefore it is important to consider the degree to which the results apply to a broader population. Third, all of these methods implicitly rely on throwing out variation in the data (the non-exogenous variation). In other words, they involve losing power in order to address exogeneity. This means that quasi-experimental work cannot use the R-squared as a useful summary of the appropriateness of the model. While R-squared or a comparison of log-likelihoods is very useful in many other contexts, it is not the focus of quasi-experimental papers (Goldfarb & Tucker, 2014).

6. What happens if there is no quasi-experiment?

Sometimes, there is no quasi-experiment in the data. The question then is whether controls are sufficient to deal with the omitted variables problem. Adding controls through multiple regression addresses potential bias in the treatment effect by including covariates that are correlated with the treatment and the outcome in a linear way. As typified by Wooldridge (2000), a realistic perspective for such an approach is that we can hope to infer causality." Caution is warranted for two reasons. First, the linear multiple regression model assumes a particular functional form. Second, and more importantly, it is not possible to know whether the controls capture all of the relevant omitted variables. Matching estimators help address the first concern. A matching estimator compares the outcomes of a treatment group and an artificial control group, where the treatment and controls groups are ‘matched’ based on similarity on observed characteristics . Thus, rather than assuming the linear structure, matching estimators allow for a non-parametric (i.e. flexible) relationship for controlling observables. If outcome measures are costly to obtain, matching saves time and effort in the data collection process. Matching estimators are often mentioned as a solution to the potential outcomes problem. This is not quite accurate. It is true that matching estimators allow for flexible controls for observables; however, matching estimators do not address the second (and more substantive) concern in using multiple regression to infer causality { that it is not possible to know whether the included covariates capture all of the relevant variables . There are a variety of different matching estimators. What they have in common is the

ability to identify similar individuals based on observables in a flexible way. Typically, this involves a 'propensity score,' which is a statistical prediction of how likely an individual is to be in the treatment group. In the absence of the quasi-experimental variation which is the focus of this guide, researchers can at least derive a measure of how large the omitted variables bias has to be in order to change the conclusions. As mentioned above, Altonji et al. (2005) provide a method to assess how big the omitted variable bias has to be relative to the included controls in order for the documented results to go away. In this way, when the most obvious confounds have little impact, it provides a limited way to assess the plausibility of the causal interpretation even in the absence of a quasi-experiment. This strategy can be enhanced by showing multiple data sets with different potential biases yield similar results. In some cases in the absence of a quasi-experiment, the uses of such techniques are sufficient. In particular, if the researchers do not have reason to expect reverse causality, if the researchers have included a large number of controls (including the most obvious confounds), and if these controls do not change the estimated treatment effect, then it is reasonable for the researchers to clearly state the assumptions behind the interpretation and move to exploring the mechanism (Goldfarb & Tucker, 2014).

7. Conclusion

Quasi-experimental studies offer important opportunities to increase and improve evidence on causal effects: It can generate causal evidence when randomized controlled trials are impossible; typically generate causal evidence with a high degree of external validity; they avoid the threats to internal validity that arise when participants in non-blinded experiments change their behavior in response to the experimental assignment to either intervention or control arm (such as compensatory rivalry or resentful demoralization); they are often well-suited to generate causal evidence on long-term outcomes of an intervention such as health, economic and social consequences; and they can often generate evidence faster and at lower cost than experiments and other intervention studies.

Quasi-experiments offer the practical advantages that they can be carried out when randomized experiments are not possible. They have the important advantages that they typically generate results that are of higher external validity than experimental results, because they take place in 'real world' settings rather than in the artificial context of experiments. They further avoid the threats to internal validity that arise when participants in non-blinded experiments change their behavior in response to the experimental assignment,

such as compensatory rivalry or resentful demoralization. Quasi-experiments are also well suited to establish causal effects on long-term health outcomes, as well as on non-health outcomes of a health intervention, such as social and economic consequences. Finally, quasi-experiments often generate results faster and at lower costs than experiments.

References

Anderson, S. (2017). *Quasi-experimental methods.* London: The International Institute for Environment and Development.

Balvin, N., Hague, S., & Jackson, D. (2014). *Quasi-Experimental Design and Methods.* UNICEF.

Banerjee, A., & Duflo, E. (2009). Annual Review of Economics. *The Experimental Approach to Development Economics* , 1:1.1-1.28.

Bärnighausen, T., Tugwell, P., Røttingen, J.-A., Shemilt, I., Rockers, P., Geldsetzer, P., . . . Brown, A. (2017). Quasi-experimental study designs series. *Journal of Clinical Epidemiology*.

Bor, J. e. (2014.). Regression discontinuity designs in epidemiology: causal inference without randomized trials. . *Epidemiology*, 25(5): p. 729-37.

Goldfarb, A., & Tucker, C. (2014). *Conducting Research with Quasi-Experiments: A Guide Marketers.*

Intrac. (2017). *Quasi- Experimental approaches.* Intrac.

Meyer, B. (1995). Natural and quasi-experiments in economics. *J Bus Econ Stat*, 151-162.

Rogers, J., & Révész, A. (2019). *Experimental and quasi-experimental designs.*

Stern, E., Stame, N., Mayne, J., Forss, K., Davies, R., & Befani, B. (2012). *Broadening the Range of Designs and Methods for Impact Evaluations.* Report of a study commissioned by the Department for International Development (DFID), Working paper.

Thompson, C. B., & Panacek, E. A. (2006). Research Study Designs: Experimental and Experimental. *Air Medical Journal*, 242-246.

Trochum, W. (2006). The Non-equivalent Groups Design. *Research Methods Knowledge Base*.

White, H., & Sabarwal, S. (2014). *Quasi-Experimental Design and Methods. Methodological briefs, impact evaluation.* UNICEF.