Beyond Baseline and Follow-up:

The Case for More T in Experiments[*]

David McKenzie, *World Bank*

Abstract

The vast majority of randomized experiments in economics rely on a single baseline and single follow-up survey. If multiple follow-ups are conducted, the reason is typically to examine the trajectory of impact effects, so that in effect only one follow-up round is being used to estimate each treatment effect of interest. While such a design is suitable for study of highly autocorrelated and relatively precisely measured outcomes in the health and education domains, this article makes the case that it is unlikely to be optimal for measuring noisy and relatively less autocorrelated outcomes such as business profits, household incomes and expenditures, and episodic health outcomes. Taking multiple measurements of such outcomes at relatively short intervals allows one to average out noise, increasing power. When the outcomes have low autocorrelation and budget is limited, it can make sense to do no baseline at all. Moreover, I show how for such outcomes, more power can be achieved with multiple follow-ups than allocating the same total sample size over a single follow-up and baseline. I also highlight the large gains in power from ANCOVA analysis rather than difference-in-differences analysis when autocorrelations are low and a baseline is taken. This article discusses the issues involved in multiple measurements, and makes recommendations for the design of experiments and related non-experimental impact evaluations.

Keywords: Randomized Experiments; Multiple Measurements; Program Evaluation.

JEL codes: O12, C93.

[*] I thank Chris Woodruff and participants at the Warwick University summer workshop on Firms in Development for conversations and questions which lead to this article; Dean Karlan (the editor), two anonymous referees, Chris Blattman, Miriam Bruhn, Jishnu Das, John Gibson and Berk Özler for helpful comments; Matthew Groh for research assistance; and Miriam Bruhn, Jishnu Das, and Markus Goldstein for sharing their data.

1. Introduction

The number of randomized experiments being conducted in economics has exploded over the past decade, especially in development economics. The vast majority of these studies use only a single baseline and single follow-up survey, or a single follow-up with no baseline. This has been the case in the early randomized experiments looking at education (e.g. Glewwe et al, 2004) and health outcomes (e.g. Miguel and Kremer, 2004), and has remained true as experiments have expanded to consider other interventions and outcomes, such as recent high-profile experiments in microfinance (Karlan and Zinman, 2011; Banerjee et al, 2010). In the rare cases where multiple post-treatment survey waves have been conducted, they have been typically taken relatively far apart in time, with the goal of examining whether the treatment effect differs in the short-term and medium-term. For example, Banerjee et al. (2007) examine impacts of educational interventions at one year and two year horizons. Indeed, so much is the paradigm of baseline plus follow-up accepted that the excellent toolkit for randomization of Duflo et al. (2008) does not discuss at all the possibility of doing more than one pre-treatment or post-treatment round of surveying, let alone the choice of how many such rounds.

In contrast, the clinical trials literature has noted the potential advantages of taking repeated measures of outcomes of interest not just to study the time course of treatment effects, but to obtain more precise estimates of effects around particular endpoints (Frison and Pocock, 1992). Vickers (2003) argues that the number of repeat measures should be a key design choice in conducting experiments, and concludes that the benefit of such additional measures is of greatest value when the autocorrelation of measures is low, such as with episodic conditions like headaches. Such a description would certainly seem to fit key economic outcomes like business profits, and incomes and consumption of the poor. As the recent work by Collins et al. (2009) makes abundantly clear, one of the key difficulties of living on $2 a day is that people don't receive $2 every day, but rather a highly irregular stream of income. Measuring microenterprise income, in particular, can be difficult, with the resulting data typically having large heterogeneity in reported profits among reasonably similarly sized firms (de Mel et al, 2009). While some of this likely reflects measurement error, Fafchamps et al. (2010) report that Ghanaian microenterprise owners confirm 85 percent or more of changes in profits above 150 percent or

below -60 percent as genuine, reflecting seasonality and the high degree of idiosyncratic variability facing microenterprise owners.

As a result of the high variability and low autocorrelation in economic outcomes like firm profits, income, and expenditure among poor households, there is much to be potentially gained by taking multiple measures of these outcomes at relatively short intervals and averaging over them when estimating treatment effects. This is the approach used in de Mel et al. (2008), who use a baseline and 8 quarterly follow-up waves of business profits and in Fafchamps et al. (2011) which uses two survey rounds before randomization and a further 4 quarterly follow-up surveys.[1] However, this approach seems to be the exception rather than the rule. The premise of this paper is that many field experiments are making a suboptimal choice of how many rounds of surveys to collect, and that the default choice of a single baseline and single follow-up study is unlikely to be optimal in many cases. This paper aims to combine insights from the medical literature on repeated measurement, new analysis of the formulae underlying power calculations, and experience from these studies using multiple measures in economics to provide a practical guide for researchers designing experiments. While our discussion will be in terms of experiments, many of the same issues and therefore lessons also apply for design of non-experimental impact evaluation designs, such as matched difference-in-difference estimation.[2]

An additional contribution of the paper is to draw to the attention of empirical researchers in economics the large improvement in power than can arise when estimating treatment effects via an Analysis of Covariance (ANCOVA) estimation compared to using the more common difference-in-difference specification. The improvement in power is greatest when the autocorrelation is low – intuitively when the baseline data have little predictive power for future

[1] Bloom et al. (2011) provides an example of very large T, with 114 weekly observations on firm output, quality, and inventory levels.

[2] Gibson and McKenzie (2010) provide an example where multiple rounds of follow-up data are averaged to get more precise measurements of consumption and income in a matched difference-in-differences impact evaluation of a seasonal migration program. One key difference with non-experimental evaluations is that the presence of a baseline becomes more valuable to allow one to control for baseline differences across individuals; whereas randomization ensures these baseline differences are balanced on average. In addition, multiple rounds of pre-treatment data are also of additional use in non-experimental studies to show parallel trends in difference-in-differences analysis, or to rule out or control for the presence of pre-program effects like an Ashenfelter dip. Multiple pre- and post-treatment data can also be combined to conduct overidentification tests to distinguish between different data generating processes (Ashenfelter and Card, 1985).

outcomes, it is inefficient to fully correct for baseline imbalances between treatment and control groups.

The bottom line of this paper is that collecting multiple measurements post-treatment will make most sense when the data have low autocorrelation and where the definition of the outcome itself doesn't change with the measurement frequency; whereas baseline data is of most use when the autocorrelation is high. I show that for many economic outcomes, experiments with a fixed budget to conduct a total of K surveys would have greater power when dividing that K over multiple post-treatment rounds than having a baseline and single follow-up. However, there are many other reasons to undertake a baseline survey, and the point of this paper is not to argue that baselines should never be taken. Instead, the paper implies that taking multiple measurements can be an important way to improve power in cases where the total cross-sectional size is limited or the intervention is expensive; and that there are likely to be a number of circumstances where taking two follow-up surveys with a smaller sample instead of one with a larger sample may make sense, and that in terms of power alone, there will be no cases where no baseline is the budget-constrained optimal choice.

Section 2 begins with a theoretical discussion of what power calculations tell us about the gain to be had from using more rounds of data, and provides examples for common types of economic data. Section 3 then derives implications for the choice of how many pre-treatment and post-treatment waves to use. Section 4 then discusses additional practical issues which are likely to arise when considering multiple measurements in economic experiments, many of which are not common concerns in clinical trials. Section 5 concludes.

2. What is the gain from more rounds?

Let $Y_{i,t}$ be an outcome of interest for household or firm i in survey round t. Consider an intervention which assigns units to receive a binary treatment (such as getting business training or not, or getting a conditional cash transfer or not). Suppose there are m pre-treatment survey rounds (labeled $-(m-1)$ through 0) and r post-treatment survey rounds (labeled 1 through r). A common method in economics for estimating the treatment effect[3] of the

[3] I focus on estimation of intention-to-treat effects here, but the implications are similar for measurement of the treatment effect on the treated and other treatment effects of interest.

intervention on the outcome of interest is via the following difference-in-differences specification:

$$Y_{i,t} = \beta EVERTREAT_i + \sum_{t=-(m-1)}^{r} \delta_t + \gamma TREAT_{i,t} + \varepsilon_{i,t} \quad (1)$$

where $EVERTREAT_i$ is a dummy variable which takes value one if unit i is assigned to the treatment group and zero if it is assigned to the control group, the δ_t are time dummies which capture the mean for the control group in each time period, and $TREAT_{i,t}$ takes value one if unit i has been assigned to receive treatment by time t (that is for t=1,2,…r), and zero otherwise. The treatment effect of interest is then given by γ. Assume the $\varepsilon_{i,t}$ are independent across individual units with cross-sectional variance σ^2, but may be autocorrelated over time for the same unit.

Note that γ does not contain a t subscript: either the treatment is assumed to have a constant level effect, or the treatment effect of interest is taken to be the average treatment effect over the r post-treatment rounds. Clearly a further, very important, use of multiple post-intervention surveys is to collect information on the trajectory of treatment impacts (e.g. Woolcock, 2009), since the short and long-run impacts of some policies may differ dramatically. However, unless one imposes some structure on the form these impact trajectories can take, collecting more rounds of data merely increases the number of points in time at which these impacts can be measured, but does not improve the power of an experiment for measuring these impacts. Even in this case where impacts at different points in time are of interest, one can consider multiple measurements located around each time horizon of interest, and thereby improve power. For example, if a study is interested in knowing whether the impact of microfinance differs at a one year horizon versus a three year horizon, taking measures of microenterprise profits at 11, 12 and 13 months after treatment, and at 35, 36, and 37 months after treatment can be used to improve the accuracy of impacts at around these horizon points, by assuming the treatment effect is constant or considering an average treatment effect in the neighborhood of the horizon of interest. The analytics and ideas of the remainder of the paper can then be applied to estimating the impact at a particular horizon by taking more measurements in neighborhoods around these horizons.

2.1.The stylized case with multiple rounds pre- or post-treatment.

Consider the general case of *m* pre-treatment survey rounds and *r* post-treatment survey rounds. The difference-in-differences estimator is obtained by estimation of equation (1). This can be written as:

$$\hat{\gamma}_{DD} = (\bar{Y}^T_{POST} - \bar{Y}^C_{POST}) - (\bar{Y}^T_{PRE} - \bar{Y}^C_{PRE}) \quad (2)$$

where for $g \epsilon \{T, C\}$

$$\bar{Y}^g_{POST} = \frac{1}{r}\sum_{t=1}^{r} \bar{Y}^g_t = \frac{1}{rn_g}\sum_{t=1}^{r}\sum_{i=1}^{n} Y^g_{i,t}$$

is the mean of the outcome for treatment or control group g in the post-treatment period, and n_g is the number of cross-sectional observations assigned to group g. The pre-treatment means $\bar{Y}^T_{PRE}$ are defined $\bar{Y}^C_{PRE}$ analogously.

Frison and Pocock (1992) give a general formula for the variance of $\hat{\gamma}_{DD}$, which depends on the full pattern of autocorrelations across the different time periods. If one knows this full covariance matrix it can be used, but in practice this full covariance structure is unlikely to be known, and power calculations are performed after making a number of simplifying assumptions. Frison and Pocock assume equal variances for all time points, equal correlations between all pairs of time points, and that autocorrelations are equal for the treatment and control groups. They show for physical health measures like cholesterol, blood pressure, and CD_4 cell count that this assumption of constant autocorrelation is reasonable. In the next subsection I show such an assumption may also hold approximately for key economic outcomes. Then assuming further that the treatment and control groups are each of cross-sectional size $n_T = n_C = n$, and denoting the constant autocorrelation by ρ, the variance of the difference-in-differences estimator is shown by Frison and Pocock to be:

$$\frac{2\sigma^2}{n}\left[\frac{1+(r-1)\rho}{r} - \frac{(m+1)\rho - 1}{m}\right] \quad (3)$$

Alternatively one can ignore the pre-treatment survey rounds and just calculate the simple difference in means post-treatment:

$$\hat{\gamma}_{POST} = (\bar{Y}^T_{POST} - \bar{Y}^C_{POST}) \quad (4)$$

Under these same assumptions, the variance of this difference in post-treatment means is then:

$$\frac{2\sigma^2}{n}\left[\frac{1+(r-1)\rho}{r}\right] \quad (5)$$

Equations (3) and (5) are respectively the variances used in power calculations under the "POST" and "CHANGE" methods reported in STATA's *sampsi* command. We therefore see that:

Implication 1*:* With *m* baselines and *r* follow-ups, difference-in-differences only gives more power than the post estimator which ignores the baseline data when the autocorrelation is greater than *1/(m+1).*

In particular, in the standard case of a single baseline and follow-up, difference-in-differences will only be preferred to the post estimator in terms of power when the autocorrelation is greater than 0.5. Intuitively, the difference in two random variables has a higher variance than that of one of these variables alone unless those variables are sufficiently highly correlated. As more baseline rounds of data are available, it becomes more costly to ignore them. However, as with the single baseline case, if the data are only weakly autocorrelated, it can be costly in terms of power to fully control for baseline differences in means via difference-in-differences.

Conditioning on a variable which is correlated with the outcome of interest can reduce the variance of the treatment estimator in a randomized trial. An important estimator which does this is the ANCOVA estimator:

$$\hat{\gamma}_{ancova} = (\bar{Y}^T_{POST} - \bar{Y}^C_{POST}) - \hat{\theta}(\bar{Y}^T_{PRE} - \bar{Y}^C_{PRE}) \quad (6)$$

This can be estimated via a least squares regression of the following equation for t=1,2,..,*r* and for individual *i*:

$$Y_{i,t} = \sum_{t=1}^{r} \delta_t + \gamma TREAT_{i,t} + \theta \bar{Y}_{i,PRE} + \varepsilon_{i,t} \quad (7)$$

where $\bar{Y}_{i,PRE}$ is the mean for individual i over the m pre − treatment rounds. It is very rare for economists to collect more than one round of data pre-treatment, so such an equation has not been typically estimated in practice for more than one pre-treatment round. Frison and Pocock (1992) show that ANCOVA is more efficient than either difference-in-differences or the post estimator, and under the same assumptions used to derive (3), has a variance of approximately[4]:

$$\frac{2\sigma^2}{n}\left[\frac{1+(r-1)\rho}{r} - \frac{m\rho^2}{1+(m-1)\rho}\right] \quad (8)$$

This is the formula used in power calculations under the ANCOVA option using STATA's sampsi command.

Implication 2*:* If the autocorrelation is zero, then only the number of post-treatment survey waves affects the power of the ANCOVA estimator, there is no gain from baseline surveys, and the variance reduces to that of the POST estimator. Intuitively, an autocorrelation of zero means that each round the outcome is a mean plus noise. The treatment changes the mean, and more post-treatment rounds enables one to better average out this noise. When the autocorrelation is low, the baseline data are still not that informative about what future values of the outcome of interest will be, and controlling fully for the baseline difference in means via difference-in-differences will therefore overcorrect for differences which don't have much predictive power. ANCOVA adjusts the degree of correction for baseline differences in means according to the degree of correlation between past and future outcomes actually observed in the data.

With a single baseline and follow-up, the ratio of the difference-in-differences variance to the ANCOVA variance is 2/(1+ ρ). So when ρ=0, with a single baseline and follow-up, one would need twice the sample size when using difference-in-differences to get the same power as obtained with ANCOVA. When ρ=0.25, which we will see to be a reasonable estimate for several economic outcomes, the sample size needed is still 60% higher with difference-in-differences than with ANCOVA to get the same power. There are therefore important gains in

[4] This is approximate since it ignores the degree of freedom involved in estimating the coefficient on the lagged outcome and the sampling error in this estimate. This correction is negligible in any reasonable sized randomized trial (see Frison and Pocock, 1992).

power to be had from not using difference-in-differences to estimate treatment effects with standard economic variables.

This comparison of the ANCOVA and difference-in-difference estimators in terms of efficiency is appropriate for experimental estimation, where both estimate the same average treatment effect. However, in non-experimental estimation, the estimators also require different assumptions for consistency (see Imbens and Wooldridge, 2009, p. 70.). As a result, even though ANCOVA should be preferred for experimental designs, I will continue to also discuss implications for difference-in-differences analysis since it may be preferred for some non-experimental designs, as well as being common practice in many experimental studies. Note further that the comparison of ANCOVA and the post estimator is based on the assumption that a baseline survey has been taken, and that the post estimator is just choosing to ignore it. We will see that once the number of surveys becomes a choice parameter, there are also cases in which the POST estimator (with say 2 follow-ups and no baseline) is preferable to the ANCOVA estimator (with say 1 follow-up and 1 baseline).

2.2 Autocorrelations in practice for economic experiments

As noted above, the formulae given for the variances of the different estimators assume that the autocorrelations are equal between all points in time, and that they are equal for the treatment and control groups. For power calculations to be useful, we need not have these assumptions hold exactly, but we need them to be a reasonable approximation. I therefore first ask whether empirically these may be reasonable assumptions for many economic applications, and then, in the next sub-section, discuss what to do in cases when this assumption doesn't hold.

The assumption that the autocorrelations are equal between all points in time is one that many economists may have second thoughts about for both theoretical and statistical reasons. On the theoretical side, if we think that a large source of the period-to-period variation in economic outcomes like profits, income or consumption comes from temporary shocks, we should expect the impact of these shocks to die out over time, so that observations closer apart in time will be more highly correlated than those further apart in time. Statistically, constant (non-zero) autocorrelations imply that although the data generating process is covariance-stationary, it is

non-ergodic for the mean and thus a law of large numbers will not apply to the mean of $\varepsilon_{i,t}$ as the time dimension approaches infinity.

However, while such an assumption may therefore not be appropriate globally over all horizons, it may be a reasonable assumption within the likely time horizons of many applications. That is, while we might not believe the correlation between business profits this month and next month is the same as that between this month and 10 years from now, the correlation between this month and next month may be close to that between this month and two months from now. For example, in a local neighborhood of a particular time horizon, a reasonable data generating process for $\varepsilon_{i,t}$ might be $\varepsilon_{i,t} = \alpha_i + u_{i,t}$, where $u_{i,t}$ is i.i.d. noise. Thus, for example, within a few months around the one year after an intervention horizon we might expect profits for a particular firm to be higher or lower depending on some longer-term shock α_i, and then in addition from month to month the firm gets idiosyncratic demand shocks which cause its profits to move around further.

Ultimately it is therefore an empirical question as to whether this assumption seems reasonable over the periods of typical economic experiments, and especially over short-term neighborhoods of a few months around a particular horizon of interest. Figures 1A, 1B, and 1C investigate this assumption for microenterprise profits and for household expenditure. Figure 1A shows autocorrelations for monthly profits from the Sri Lankan microenterprises in de Mel et al. (2008) which were measured at three month intervals, and uses data from similar microenterprises measured monthly in de Mel et al. (2009) to show the autocorrelation in monthly profits at the one month and two month horizons.[5] A horizontal line is drawn at 0.38, which is the average autocorrelation over the different interval measures. Pointwise confidence intervals for these autocorrelations are obtained via 1000 bootstrap replications. While Figure 1A does suggest some tendency for the autocorrelation to fall as the time horizon increases, the data do not deviate much from the 0.38 level over all time horizons between 1 month and 18 months, and so the assumption of constant autocorrelation seems a reasonable approximation here – with the next subsection showing that using this average level will not lead to very different choices than taking full account of the correlation structure for deviations from equality of this magnitude.

[5] These autocorrelations use the raw data, reflecting what might be used for estimating treatment effects. Nevertheless, the low autocorrelations are not driven by outliers, with the Spearman (rank-order) autocorrelations also all below 0.5.

Figure 1B shows the autocorrelations in monthly microenterprise profits in urban Ghana, which were measured at three month intervals in the experiment of Fafchamps et al. (2011), and Figure 1C shows the autocorrelation in three monthly household expenditures from the same experiment, also measured at quarterly intervals. Horizontal lines are drawn at 0.32, the mean autocorrelation over these horizons for profits, and 0.20, the mean autocorrelation over these horizons for household expenditures. Again both figures suggest the assumption of constant autocorrelation holds approximately over these horizons.[6] Gibson et al. (2003) also provide evidence from Zambia and urban China that the autocorrelation of expenditure does not change very much within a year as one increases the number of months between surveys.

Table 1 then considers the evidence for the second part of the assumption, which is that the autocorrelations are the same for treatment and control, using the microenterprise profits data from Sri Lanka and Ghana. The point estimates in most periods are reasonably similar between treatment and control, and the bootstrapped confidence intervals for the difference in autocorrelation between the two always contain zero. Evidence from another setting comes from using the test score data of Banerjee et al. (2007). The verbal test score autocorrelation is 0.58 for the treatment group and 0.61 for the control group; and the math test score autocorrelation is 0.47 for the treatment group and 0.55 for the control group. Therefore in these samples this assumption seems reasonable.

Table 2 provides autocorrelations from other data sets for a range of different economic outcomes. The purpose of this is twofold. First, it demonstrates that for many economic outcomes, the autocorrelations are typically lower than 0.5, with many around 0.3. An exception is test scores, which have higher autocorrelations. Second, it provides some parameters that researchers can use when conducting their own power calculations, since it is rare to have autocorrelation data available for the study populations being considered. We see from Table 2 that autocorrelations are often in the 0.2-0.3 range for household income and consumption. In rural Ghana, the autocorrelation in Boozer et al (2010)'s data is 0.32 when wives report on both their own and their husband's consumption, but increase to 0.58 to 0.66 when husbands and wives report separately on consumption and the results are combined. The autocorrelations are

[6] The Sri Lankan microenterprise surveys only measured household expenditure annually, which is why household expenditure is not shown for this sample.

lower for the poor – in both Mexico and Argentina the autocorrelation of labor income is around 0.2-0.3 for individuals with first period labor income below the median, compared to autocorrelations of 0.5-0.8 for those with this income above the median. This is consistent with the evidence in Collins et al. (2009) that incomes for the poor are very volatile.

The autocorrelation will typically be lower in the more homogeneous samples typical of many randomized experiments than it is in general samples of the population. It will also be lower if one stratifies the randomization or uses fixed effects and thereby controls for time-invariant reasons why one household or firm may have higher income, expenditure or profit levels than another. To see this in a simple example, write the error term $\varepsilon_{i,t} = \theta X_i + v_{i,t}$, where X is a characteristic that is used to stratify the randomization on, or to restrict an experiment only to individuals with certain values of X. Then:

$$Var(\varepsilon_{i,t}) = \theta^2 Var(X_i) + Var(v_{i,t})$$

and

$$Cov(\varepsilon_{i,t}, \varepsilon_{i,t-s}) = \theta^2 Var(X_i) + Cov(v_{i,t}, v_{i,t-s})$$

Controlling for X via stratified randomization, or restricting the sample only to those with certain values of X will therefore reduce the variance of $\varepsilon_{i,t}$, which increases the power of an experiment. However, it also reduces the autocorrelation (by removing a factor which would otherwise cause the error term to be correlated from one period to the next), thereby increasing the value of more measurements.

2.3 What should one do if the autocorrelations are not constant?

The empirical evidence therefore suggests that the assumption that the autocorrelation is constant over different time horizons and is the same for the treatment and control is likely to be a reasonable approximation in a number of economic experiments involving continuous and somewhat noisy outcomes like profits, income and consumption. How do violations from this assumption affect power calculations?

First consider violations of the assumption that the autocorrelation is the same in the treatment and the control groups. For example, Drexler et al. (2010) and Karlan and Valdivia (2011) find

evidence that business training programs improve sales in bad months, without any significant change in mean sales. The training might cause sales to be less sensitive to external conditions, thereby increasing the autocorrelation. Then in this case, one can show that for a given total cross-sectional sample size, for estimation using the post estimator, one should choose the treatment group sample size n_T and control group sample size n_C such that:

$$\frac{n_T}{n_C} = \sqrt{\frac{1+(r-1)\rho_T}{1+(r-1)\rho_C}} \qquad (9)$$

where ρ_T and ρ_C are the autocorrelations for the treatment and control groups respectively. Intuitively more time series observations add less information when the data are highly autocorrelated, so it is optimal to allocate more cross-sectional sample to the group for which the time series data is less informative. Panel A of Table 3 show what this implies in practice – for example, when the treatment group has autocorrelation of 0.7 and the control group 0.5, with three follow-up surveys the treatment group should be 1.095 times the size of the control group. That is, rather than a 50:50 split of the sample, it should be split 52:48. One sees that even with 5 follow-ups and a massive difference in autocorrelations (0.2 vs 0.8), the optimal split is still 40:60, or not that far from 50:50. In most practical applications then, the types of deviations from the assumption of equal autocorrelation that are likely to arise are not going to imply much difference in the allocation of treatment and control sample sizes.

Similarly, by the symmetry of the variance components, the optimal allocation of treatment to control should be the square root of the ratio of the expression in equation (3) using the treatment group parameters to this expression using the control group parameters. When there is only one pre-treatment survey round, this simplifies to:

$$\frac{n_T}{n_C} = \sqrt{\frac{1-\rho_T}{1-\rho_C}} \qquad (10)$$

which we see does not depend on the number of follow-up survey rounds. Intuitively difference-in-differences performs badly when the autocorrelation is low, and so optimally one needs to have relatively more sample from the group with the lowest autocorrelation. Panel B of Table 3 shows how this plays out in practice. Again the ratio of the treatment to the control group sample

size is not very different from one for differences in the autocorrelations in the 0.2 to 0.3 range. However, for larger differences the ratio grows.

Finally, for ANCOVA the optimal allocation should be the square root of the ratio of the expression in equation (8) for the treatment group to that for the control group. Here the tendency to put relatively more sample in the group with a higher autocorrelation for which more T is less informative clashes with the tendency to put relatively more sample in the group with a lower autocorrelation for which differencing is less effective. The result, shown in panel C of Table 3, is that the optimal sizes of treatment and control groups doesn't deviate that far from a 50:50 division. Thus deviations from the assumption of equal correlations for the treatment and control groups shouldn't matter much for power calculations in most practical applications.

Consider next the assumption that the autocorrelations be constant over the different time periods in the study. What should one do if this is violated? For post and difference-in-differences analysis, the variances are linear in the autocorrelations, and so power calculations using the mean ρ will still give the correct power for these estimators even when the autocorrelation varies over different time periods. So, for example, if the error term follows a AR(1) process such that ρ is 0.8 over 3 months, $0.8^2 = 0.64$ over 6 months, $0.8^3 = 0.512$ over 9 months, and $0.8^4 = 0.4096$ over 12 months, one can simply take the mean of the off-diagonal elements of the correlation matrix, which here would be 0.655 for five rounds of surveys, and use this in the power calculations. Figure 1 shows in practice the simple mean autocorrelation is not that different from the round by round autocorrelations in some practical cases.

In contrast, since the variance of the ANCOVA estimate is decreasing in ρ^2, using the average ρ will understate the sample size one would actually need to achieve a given power. In such cases using the minimum ρ expected among follow-up rounds would be conservative. Of course in the unlikely event that one actually knows the exact structure the autocorrelations will follow, the exact formula could instead be used.

2.4 Where do the gains from more T come from?

A somewhat subtle, but crucial, assumption in the above discussion is that the definition of the outcome measure does not change when we change the number of time periods. This is natural for the types of measures used in the medical fields, where blood pressure, CD4 counts, and

cholesterol are all instantaneous snapshot measures. Economic measures which are stocks (like asset holdings) or which are also snapshot measures (like test scores) likewise do not change definition with the frequency of measurement. In contrast, a number of outcomes in economic experiments are flow measures, such as income, expenditure, or business profits over a given time period. The above analysis then shows that there is more power to detect an impact of a given treatment on monthly profits if we measure monthly profits every month for three months, rather than just do a single follow-up survey and measure profits in the last month.

What the above analysis does *not* imply is any preference for using, for example, three survey waves of monthly profit data versus one survey round in which profits are measured over the last three months. Assuming a constant treatment effect and no measurement error, the power to detect this treatment effect will instead be the same if we aggregate up the three months of profit and estimate the impact on three month profits, as if we use the three rounds of profit data as panel data in estimating (2), (4) or (6). Likewise, we can obtain a much more accurate measure of the impact of an intervention on daily profits by measuring 30 days of profits and using them in the estimation than if we just looked at the impact on the last day's profits, but don't do better from measuring the impact on daily profits from 30 days of measurement than if we looked at the impact on the last months profits.

Where the gains from multiple measurements come from with flow variables is therefore in the ability to either extend the time horizon which measurement takes place, or to reduce measurement error within this horizon.[7] It is common practice for surveys to only ask about the last month's profits or the last week's food expenditure, because microenterprise owners and households may not be able to recall data accurately for longer periods. So having multiple survey rounds, each of which asks about the flow variable over this pre-specified period is better than having just one round which does the same. Alternatively, one might choose to push firm owners or households to recall data over longer horizons, but there is likely to be more measurement error in doing so, and in such cases multiple measurements can help improve power by reducing the noise in this aggregate.

3. Implications for choosing the number of pre-treatment and post-treatment rounds

[7] When treatment effects vary over time, multiple time periods offers further advantages in providing more degrees of freedom to estimate the trajectory of impacts.

Based on the previous section, the stylized case in which autocorrelations are assumed to be the same for treatment and control groups and over multiple periods appears a reasonable approximation for many practical cases, with the types of deviations from these assumptions found in practice not greatly affecting the use of these formulae. Equations (3), (5) and (8) can be used to help researchers decide how many rounds of data they should collect before and after the treatment. We use these formulae to examine implications for choice of the number of pre-treatment and post-treatment rounds.

3.1. Given a fixed T of 3 or more rounds, how should they be split between pre-treatment and post-treatment?

The first question we ask is how a researcher who has decided on fielding *T* multiple survey rounds should split these rounds before and after treatment. To do this, we solve for the choice of *r* and *m* which minimizes the variance given *r+m=T*.

Implication 3: The optimal choice is *r = T/2 = m* for difference-in-differences estimation, and this choice does *not* depend on the autocorrelation in the data. When $T \geq 3$ is odd, the power is the same when choosing *m-r=1 as r-m=1*, that is, when allocating the extra odd wave to pre-treatment or to post-treatment.

Proof: Setting *r+m=T* in equation (3) and solving for the optimal choice of *r* to minimize the variance gives the first part of this result. Re-writing the variance in (3) as

$$\frac{2\sigma^2}{n}\left[\frac{(r+m)(1-\rho)}{rm}\right] \qquad (11)$$

shows that *r* and *m* contribute symmetrically to the variance of the difference-in-differences estimator, and so that choosing *r = (T+1)/2* and *m=(T-1)/2* will yield the same results as choosing *r=(T-1)/2* and *m=(T+1)/2*

Intuitively, in a difference-in-differences design, differences in pre-treatment means have exactly as much weight in the estimator as differences in post-treatment means, and so the optimal strategy is to have the same number of time periods to estimate the difference pre-treatment as post-treatment. Note from (11) that while the autocorrelation does not affect the choice of how to

allocate the survey rounds between pre-treatment and post-treatment in a difference-in-differences design, it does affect the cross-sectional sample size needed – a higher autocorrelation reduces the variance, thereby requiring a smaller *n*.

Implication 4: For ANCOVA estimation, the optimal choice of *r* given fixed *T* is given by

$$r = \frac{1 + \rho(T-1)}{2\rho} \qquad (12)$$

So when ρ=0.5, the optimal choice is *r = (T+1)/2*, whereas when ρ=0.25, the optimal choice of *r = (T+3)/2*. The nearest integer should be chosen when (12) is not an exact integer.

Proof: Expanding out the variance in (8), one sees that *r* and *m* are symmetric in the numerator. Minimizing the variance therefore amounts to maximizing the denominator, *r[1+(m-1)* ρ], which can be solved to give the expression in (12).

Thus with ANCOVA estimation, one chooses fewer pre-treatment survey rounds the lower is the autocorrelation. Intuitively, we see in equation (6) that the pre-treatment difference in means contributes less to the estimator than the post-treatment difference in means, and so it is less important to accurately measure the pre-treatment difference than to accurately measure the post-treatment difference. When the autocorrelation is low, this can result in choosing to have no baseline. For example, when T=3 and ρ=0.25, the optimal choice is *r=3*, so one is better to have 3 follow-up waves and no baseline than a baseline and two follow-up waves.

3.2 Given a fixed total sample size 2nT, what is the trade-off between a larger cross-section and more survey rounds?

In many cases a researcher will face a fixed total budget, which can fund *K=2nT* total surveys. They must then decide between carrying out a larger cross-sectional sample, which gives information on more units, and carrying out more survey rounds, which gives more information on each unit. Taking *K* as fixed, we derive the optimal division into more rounds versus more individuals per round.

Implication 5: For POST estimation, the optimal choice is *T=1* and *n=K/2* when ρ>0, and all values of *T* and *n* such that 2*nT=K* yield the same power when ρ=0. When ρ<0, the power is

higher using *T=2* and *n=K/4* than it is using *T=1* and *n=K/2*, while the assumption of constant autocorrelation across different points in time doesn't make sense for negative autocorrelation and *T>2*.

Proof: From equation (5) we can see that when $2rn=K$, the variance depends on $(r-1)\rho$, so the optimal choice of r to minimize this will depend on ρ in the way stated.

This fact is analogous to the decision in a clustered randomized trial of whether to collect more observations per cluster, or more clusters. With clustered randomization, since randomization is at the level of the cluster, if there is positive intra-cluster correlation, there is more power to be had in randomizing over more clusters than in having fewer clusters and a larger sample per cluster. Likewise here the randomization occurs in the cross-section, and so with positive autocorrelation, there is more gain to having more cross-sectional units than more observations in the time dimension on each unit.

Implication 6: For difference-in-differences estimation, the choice of how to allocate a fixed sample *K* between *n* and *T* does not affect the variance provided that one then chooses to allocate *T* equally between pre-treatment and post-treatment rounds, and that there is at least one baseline and one follow-up to enable difference-in-differences estimation.

Proof: Setting $r=m=T/2$ and $2nT=K$ in (3) yields a variance of $16\sigma^2(1-\rho)/K$ which does not depend on either *T* or *n*.

Thus the power of difference-in-differences estimation is the same with a single baseline and follow-up and *n=100*, as it is with two pre-treatment rounds and two post-treatment rounds and *n=50*. Intuitively, with difference-in-differences, every observation counts equally in calculating the pre-treatment and post-treatment means which form the basis of estimation.

This result of course depends on the assumption of equal correlation being a reasonable one. So this implication is likely to make sense when choosing between one or two follow-up surveys, or even between one and five follow-up surveys, but should not be taken to the extreme of saying we should expect the same power from *T=100* and *n=2* as we would from *T=2* and *n=100*.

Implication 7: With ANCOVA estimation, when the autocorrelation is high it is better to do a larger cross-section and fewer survey rounds, whereas when the autocorrelation is low, it is

better to do relatively more survey rounds and a smaller cross-section. Holding nT fixed, with a single baseline: (i) the optimal value of *r* holding *nr* fixed is the nearest integer to $1/\sqrt{\rho}$; (ii) two follow-up surveys will offer more power than a single follow-up survey for $\rho<0.5$; (iii) three follow-up surveys will offer more power than one or two follow-up surveys for $\rho<1/6$.

Proof: Set *m=1* and *2nr=K* in (8), and then take the derivative with respect to *r* to show (i). Compare the variances with r=1, 2, and 3, to show (ii) and (iii).

Therefore when the total sample size is fixed at *K*, power will typically be greater using a baseline and 2 post-treatment surveys with a cross-sectional sample of K/3 than a single baseline and follow-up for many types of economic data. But it will be rare in practice for more than two post-treatment surveys to be optimal with a fixed total sample size. However, the standard practice of a single baseline and single follow-up used in so many economics experiments is unlikely to be the optimal choice for all.

In practice the cost of adding another cross-sectional unit versus surveying the same unit a second time can differ. In many experiments there can be large fixed costs of adding additional units to the experiment. These fixed costs can include the costs of screening and enrolling more subjects in the study, the cost of the intervention itself (many times this is by far the largest cost), and potentially the cost of more handheld units for surveying them. Denote the fixed costs of another cross-sectional unit by *a*. There can also be fixed costs of conducting another survey round, such as the need to pay field managers for more months of work. Denote the fixed cost of another time round by *b*. Finally denote the marginal cost of another survey, whether cross-sectional or temporal, by *c*. Then, assuming only one baseline, with a fixed budget *D,* one can solve for the optimal number of follow-up rounds and cross-sectional units by minimizing the ANCOVA variance in (8) subject to $2an+br+2cn(r+1) = D$.[8] This does not have a closed form solution, but can easily be solved numerically for specific cases, and Stata code to solve this is included as an online appendix to this paper. For example, take $\rho=0.30$, b = \$100 and c = \$50, and D = \$500,000. Then with a relatively cheap treatment of a = \$100, the optimal choice is to take *n=811* and *r=3* follow-ups, whereas with an expensive treatment of a = \$10,000, the

[8] This assumes that the treatment and control groups are the same size. When the cost of adding another treatment greatly exceeds that of adding another control, optimal power with a given budget can be achieved by allocating more units to the control group than to the treatment group (See Duflo et al, 2008). Conditional on doing this, it may still be optimal to allocate fewer cross-sectional units to more survey rounds.

optimal choice is to take *n=22* and *r=25* follow-ups. Thus studies with more expensive treatments should optimally collect more data on each unit treated.

3.3. What is the marginal gain to another post-treatment round?

Finally, in many experiments researchers have a fixed cross-sectional sample, often dictated by the pool of eligible applicants for some pilot program they are investigating. For example, an NGO has money to offer a treatment to 1000 farmers, or a Government wishes to pilot a program in 50 villages. The question then facing the researcher is how many survey waves to collect. Consider the difference-in-differences estimator, and the gain in power from collecting *r+1* rounds compared to *r* rounds of post-treatment data. This gain in power is proportional to the difference in the variances, and so using (11) we have:

$$\text{Difference-in-difference Gain} = \frac{(1-\rho)}{r(r+1)} \qquad (13)$$

Note that the gain in power from adding more post-treatment rounds to difference-in-differences is the same as the gain to adding more post-treatment rounds to estimation by the post method, and thus the gain in (13) is the same as that calculated by Vickers (2003) for the case of estimation by the POST method. Using (8), we see that this is also the gain in power from adding more post-treatment rounds to ANCOVA estimation.

Implication 8: There are diminishing marginal returns to adding more post-treatment rounds in terms of the gain in power they give. The greatest gain is from moving from one to two post-treatment rounds, and the gains are smaller the higher is the autocorrelation.

Proof: One can easily determine that the derivatives of (13) with respect to both *r* and ρ are both negative.

Given the symmetry of the role of pre-treatment and post-treatment surveys in difference-in-difference estimation, we see the same holds for adding more pre-treatment rounds. For ANCOVA estimation, the gain from adding pre-treatment rounds is different from that from adding post-treatment rounds in general, and the gain from moving from one to two pre-treatment rounds is less than that from moving from one to two post-treatment rounds when $\rho < 1$.

3.4 Numerical Illustrations

To illustrate how much the required cross-sectional sample size can be reduced when more survey rounds are undertaken, Table 4 presents the minimum treatment group size *n* obtained under power calculations with different values of *r*, *m*, and ρ. I consider a hypothetical experiment intended to detect a 10 percent increase in business profits from a baseline value of 100. Microenterprise profits are noisy, so it is common for the standard deviation to be of the same order of magnitude as the mean. I therefore take the standard deviation as 100, and fix the power at 80 percent and size at 5%. Panel A then shows the treatment group sample sizes required for POST estimation, Panel B for difference-in-differences estimation, and Panel C for ANCOVA estimation, for $\rho \in \{0, 0.25, 0.5, 0.7, 0.95\}$. It is assumed that the control group is the same size as the treatment group in these calculations.

Table 4 illustrates numerically Implications 1-4, and Implication 8. Comparing the sample sizes in Panel A and B, we see Implication 1 and 3. When there is only one baseline, difference-in-differences requires larger sample sizes than post analysis when $\rho<0.5$, the same sample sizes for $\rho=0.5$, and lower sample sizes for $\rho>0.5$. The difference in sample size can be large. For example, with a single baseline and follow-up, when $\rho=0.25$, one would need a treatment group size of 2355 with difference-in-differences, compared to 1570 with POST analysis. Comparing Panels A and B to Panel C we see consistent with Implication 2 that ANCOVA requires the same sample size as POST when $\rho=0$, and otherwise requires lower sample sizes than either difference-in-differences or POST. Thus in the case of a single baseline and follow-up and $\rho=0.25$, ANCOVA requires a treatment group size of 1472, compared to the 1570 with POST and 2355 with difference-in-differences. We see in particular that ANCOVA does substantially better than difference-in-differences when ρ is low, and substantially better than POST when ρ is high.

Panel B also illustrates Implication 3. For a fixed number of waves T, the required sample size is smallest when the number of pre-treatment and post-treatment waves are equal. Thus with T=4, the treatment sample size required with $\rho=0.5$ is 785 with 2 pre-treatment waves and 2 follow-ups, compared to 982 with one baseline and three follow-ups. We also see the sample sizes required when *m=1* and *r=2* are the same as those with *m=2* and *r=1*. In contrast, as per Implication 4, we see ANCOVA favors more post-treatment waves than pre-treatment waves

when ρ is low. Thus the treatment sample size required with ρ=0.5 is 785 with one baseline and 2 follow-ups, compared to 1047 with two baselines and 1 follow-up.

Finally in Table 4 one can also observe Implication 8, the diminishing gain from adding more survey rounds, with the amount of the gain lower the higher is ρ. Thus for difference-in-differences with one baseline, when ρ=0 the gain from going from one to two post-treatment waves is a reduction in the treatment group size of 785, whereas the gain in going from two to three post-treatment waves is only 261. When ρ=0.95, the gains are only a reduction in treatment size of 39 when going from one to two post-treatment waves, and 13 in going from two to three post-treatment waves.

Table 5 illustrates the differences in power obtained in changing the allocation of a fixed total sample size between *n* and *T*. We fix $nT = 1000$, and consider different combinations of cross-sectional and time series samples. Again we assume a mean and standard deviation of 100, and a treatment effect of 10 percent. Panel A illustrates Implication 5, showing that T=1 is always best when ρ>0 for POST analysis. Panel B illustrates Implication 6, demonstrating that the power is the same regardless of how *n* and *T* are split, so long as $r=m=T/2$. Finally, Panel C illustrates Implication 7, that for a given total sample, sometimes power can be greater when doing a baseline and multiple follow-ups, than by doing a single baseline and single follow-up. For example, when ρ=0.25, power is greater doing one baseline and two follow-ups with $n=333$, than doing a single baseline and single follow-up with $n=500$, but also greater than doing one baseline and three follow-ups with $n=250$.

As a final illustration, consider the power to detect an impact of microfinance on business profits. Banerjee et al. (2010) report a mean of 550 and standard deviation of 46604 for business profits. With such noisy data, the treatment and controls group sample sizes required to detect a 10% increase in profits with 90 percent power using a single baseline and follow-up are over 15 million! Suppose then that they were able to consider a more homogeneous set of firms and measure profits more accurately[9], reducing the standard deviation to 550. Then assuming an autocorrelation of profits of 0.25, with their sample of approximately 1150 treatment and 1150 control, the power for detecting a 10% increase in profits would be 0.669 with a single cross-

[9] They construct profits as revenue less expenses. De Mel et al. (2009) show that this leads to considerably noisier profit measures than directly asking for a single profits number in their experiments.

section post-treatment, 0.697 with a baseline and single follow-up, 0.892 for a baseline and two follow-ups, and 0.952 with a baseline and three follow-ups.[10] The choice of a second or third post-treatment measure would thus allow them to detect a treatment effect of interest with considerably more power than possible with only a single survey round.

4. Recommendations for Practice and other Practical issues

The above analysis has shown that there are gains to be had from moving beyond the paradigm of single baseline and follow-up. In this section I discuss first some basic guidelines for using these results in practice, and then discuss other practical issues that might affect the choice of how many survey waves to carry out.

4.1. Guidelines for practice

1. *Highly autocorrelated outcomes:* For outcome measures like anthropometric measures or test scores, for which the autocorrelation is high (e.g. ρ=0.6 to 0.8), always include at least one baseline. Difference-in-differences and ANCOVA have much greater power than POST in these cases. Moreover, Bruhn and McKenzie (2009) also show that the power improvements from stratified or matched randomization are highest when ρ is high. However, if the total sample size is fixed, a single baseline and follow-up will offer more power than multiple pre-treatment or post-treatment rounds when ANCOVA is used. If the treatment and control sizes are fixed, the gains from going from one to two post-treatment surveys can still be non-trivial in these cases, but there is little gain from more than say 3 post-treatment surveys.

2. Outcomes with low autocorrelation: For outcome measures like business profits, incomes, or expenditure, for which the autocorrelation is typically low (e.g. ρ=0.20 to 0.40) , it can be optimal to have no baseline at all, and just do a single follow-up survey if the total sample size is limited. If the treatment group size is fixed, researchers can dramatically increase power by doing multiple post-treatment surveys. Even if a baseline is taken, researchers should not use difference-in-differences in such cases, since doing so has much lower power than either POST or ANCOVA analysis.

[10] These power calculations ignore intra-cluster correlation for simplicity, and because it is typically low for an outcome like profits.

4.2 Other Practical issues

In practice there are several other factors that should guide researchers in choosing how many survey rounds to conduct. I note some of the key factors here, and their implications for choice of the number of pre-treatment and post-treatment waves:

1. *Over what horizon is a pooled treatment effect relevant and of interest?* The analysis in this paper has assumed a common treatment effect γ. Of course if the treatment effects vary over individuals, researchers are usually content to estimate an average treatment effect, and the analysis here will still hold. However, in some experiments the effect of the treatment may also vary over time. In this case estimation of γ by pooling multiple post-treatment survey waves will yield an average treatment, where the average is over multiple waves. If the survey waves are relatively close together in time (e.g. monthly or quarterly), then combining several post-treatment waves and getting the average treatment effect over a period 9-15 months post-treatment is likely to be reasonable.[11] High seasonality may provide another reason to conduct multiple measures – it may be of more interest to get the average impact over high and low demand periods than to get the impact only at one particular point in time. In contrast, if there is reason to believe the effects differ dramatically with time since treatment, one will not want to average over long periods. Likewise, multiple measurements are sometimes used to try and understand the mechanisms through which a treatment operates, and if this is the goal, there needs to be appropriate power for estimation at a point in time, not just for the time-pooled effect. One can still boost power by taking multiple measurements in neighborhoods of these time horizons even in these cases.
2. *Implications for external validity:* The power calculations presented here concern the power to detect treatment effects in a given experimental sample. However, as the cross-sectional sample size shrinks, there is greater sampling variability, which will mean more uncertainty (wider confidence intervals) in moving from sample treatment effects to population treatment effects. As the cross-sectional sample size becomes very small, researchers will need to take care in both sampling and in comparison to broader samples

[11] Researchers interested in time since treatment effects can then combine multiple measures around 1 year, around 2 years, etc. See De Mel et al. (2008) for an example, in which quarterly waves are combined to look at how treatment effects vary over the first year since treatment, second year since treatment, etc.

to able to argue that their experimental sample is representative of a larger population of interest.

3. *Multiple outcomes may involve trade-offs:* Some studies may wish to consider impacts on both relatively highly autocorrelated outcomes (e.g. physical capital stock in a business) and relatively weakly autocorrelated outcomes (e.g. business profits). The optimal allocation of a fixed budget to cross-sectional versus time observations may differ for each, and so researchers may need to trade-off power to detect impacts on one outcome against another.
4. *Survey compliance and attrition:* Conducting multiple survey waves increases the burden on respondents, which might increase attrition. If the survey is just collecting measurement of a few key outcomes, these additional rounds of post-treatment surveys can be short and less burdensome for respondents, which can minimize this drop-out. Secondly, by conducting multiple post-treatment rounds, researchers stand a better chance of capturing at least once post-treatment individuals who are harder to track down. For example, a business owner who travels temporarily to other towns might get missed in a one-off survey, but may get found when going back at monthly or quarterly intervals. Finally, there is also the possibility that waiting for a long time to re-contact people may make it harder to find them, and make respondents think you have lost interest in them, whereas regular follow-ups may instead reduce these risks.
5. *Will multiple measures change people's reporting or people's behaviors?* Experience with measuring microenterprise profits suggests that respondents may report more accurately after one or two rounds of surveys, perhaps as they better understand the concepts being asked and can recall them better when asked for a second or third time (Samphantharak and Townsend, 2009; Fafchamps et al, 2010). This suggests a benefit to researchers of conducting more than one pre-treatment survey. However, in some circumstances there will be a concern that asking the same question multiple times will either change the way people respond, or their behavior. For example, Zwane et al. (2011) summarize the results of five experiments in which the frequency of surveying was varied. They find being surveyed more frequently lowers reported child diarrhea rates and leads to more use water treatment products and take-up of medical insurance. However, they find no effect of being surveyed on borrowing behavior. The risk is

therefore that in some settings and for some outcomes, repeated surveying may serve as reminders to increase the salience of neglected actions. Finally, repeatedly asking individuals the same questions may change reporting due to fatigue. For example, individuals who are repeatedly asked about expenditure may quickly learn that they can reduce the length of the questionnaire by saying "no" to questions which when answered positively lead to lengthy detailed questions about what was spent.

6. *Cost considerations:* Surveying a smaller cross-section over multiple survey waves can result in a different cost than doing a larger cross-section single baseline and follow-up. Whether the cost is higher or lower will depend on country context. On one hand, a smaller cross-sectional sample with more survey waves allows a smaller survey team to be used for longer periods of time, potentially resulting in higher average quality enumerators. In addition, it can lower the costs of listing since a smaller cross-sectional sample is used. Of greatest cost savings in many experiments is that it also avoids the costs of paying for the intervention for as many units. On the other hand, the team is in the field for a longer period of time, which can raise costs. As a general rule, the higher the per-unit treatment cost, the more likely it is that doing multiple measures per unit treated is optimal.
7. *Learning by doing.* Even in cases where researchers have some power to detect a treatment effect with one post-treatment survey round, it can be useful to plan multiple survey rounds relatively close together in time. This allows researchers to quickly analyze the early results from the first follow-up survey, and based on these, ask follow-up questions in the next post-treatment survey to explore new hypotheses generated by the results.
8. *How do we know what ρ is?* Power calculations already require substantial insights from researchers in terms of what the likely mean, standard deviation, and treatment effect size of interest are. Pilot surveys or existing data are often used to give some sense of likely parameters. In most cases researchers will not have data on the autocorrelation of the outcome of interest. It is hoped that the data provided in this paper for different outcomes from different surveys will therefore provide a starting point for researchers, and that more researchers will report these autocorrelations in future experiments.

9. *Aggregation and recall:* The gain in power from multiple measurements occurs when the outcome definition doesn't change with the frequency of measurement. A cheaper alternative to conducting multiple post-treatment or pre-treatment rounds is to ask multiple measurements or for aggregate measurement over a longer horizon in a single survey, with recall. While intuitively appealing, the value of this approach is likely to be limited in practice, since it is precisely the types of variables that are noisiest and hardest to measure in practice for which the data are typically less autocorrelated and for which multiple measures are of most value. For small informal firm owners who do not keep books, asking profits with recall over multiple months is difficult, while asking expenditure month by month or quarter by quarter with recall is likely to be highly inaccurate. Gibson and Kim (2010) show even with wage workers in the United States that retrospective recall is problematic, with workers underreporting transitory variation in earnings, creating non-classical measurement errors. However, such a strategy could be used in cases where good records are available, such as in experiments with larger firms which have accurate books or records.[12]

10. ***T****o baseline or not?* The power calculations above suggest that, in a number of cases, it can be optimal in terms of power to conduct a single post-treatment survey of size $n=K$ than to do a baseline and follow-up with $n=K/2$. However, there are several other factors to consider in choosing whether or not to use a baseline. A baseline can be used to stratify the randomization on key variables, improving power and providing a basis for examination of treatment effect heterogeneity. It is often used to examine what determines take-up of some intervention. It can be useful for verification of randomization in cases where there is a risk of the randomization not being implemented perfectly, and can also be used to test whether attrition is non-random in terms of baseline characteristics. Finally a baseline, and especially multiple pre-treatment surveys, offer the possibility of matched difference-in-differences as a back-up evaluation strategy in cases where there is a risk that the randomization may not get implemented in practice or a risk that take-up of a program may be lower than anticipated.

[12] See e.g. Bloom et al. (2011) who collect daily and weekly production data from plants with over 100 workers each once a month.

However, when the outcome has low autocorrelation, the baseline does not reveal very much about likely future outcomes, and thus can be an expensive undertaking. Researchers attempting to implement large-scale evaluations with Governments or NGOs, where there is reasonable uncertainty as to whether randomization will actually be carried out and whether they will have sufficient take-up. For example, consider a micro-savings intervention, where a Government or NGO agree to randomly choose villages in which to introduce a new savings product. There is a risk that after doing baseline surveys in the treatment and control villages that the Government changes its mind, and decides to introduce the product in all villages (or decides not to introduce the product at all). There is also a risk that no one uses the new product, in which case there is no possibility of measuring the impact of savings accounts on household outcomes. In such a case, practical considerations may suggest it is optimal to not carry out a baseline, and then wait and only carry out post-treatment surveys if the product is indeed randomly introduced and take-up is high.

5. Conclusions

This paper has shown that when the autocorrelation in outcome data is low, as is common with outcomes of interest like business profits, expenditure, and income, the standard paradigm of single baseline and follow-up, followed by difference-in-differences analysis, is unlikely to be optimal. Large improvements in power can be obtained from multiple post-treatment measures in experiments with fixed treatment and control group sizes, and from using ANCOVA instead of difference-in-differences. Researchers choosing how to allocate a fixed budget over multiple surveys may find they can obtain more power by not conducting a baseline at all, and if they use a baseline, will often get more power doing two follow-up waves with a smaller cross-sectional sample size than a single follow-up with a larger cross-sectional sample.

These findings are particularly likely to be of interest and use to researchers conducting experiments with interventions to help the poor, since the profits, incomes, and expenditures of the poor are typically more volatile and less autocorrelated than those of stable wage earners for example. In many cases the size of the treatment group is determined by the number of units

eligible for some pilot initiative, and so researchers can extract much more out of these samples by considering multiple measurement.

References

Andrabi, Tahir, Jishnu Das, Asim I. Khwaja, and Tristan Zajonc. (2011). "Do value-Added estimates add value? Accounting for learning dynamics." *American Economic Journal: Applied Economics*, 3(3): 29-54.

Ashenfelter, Orley and David Card (1985) "Using the longitudinal structure of earnings to estimate the effect of training programs", *Review of Economics and Statistics* 67(4): 648-660.

Banerjee, Abhijit, Shawn Cole, Esther Duflo and Leigh Linden (2007) "Remedying education: Evidence from two randomized experiments in India", *Quarterly Journal of Economics* 122(3): 1235-64.

Banerjee, Abhijit, Esther Duflo, Rachel Glennester and Cynthia Kinnan (2010) "The miracle of microfinance? Evidence from a randomized evaluation", BREAD Working Paper no. 278.

Bloom, Nick, Benn Eifert, Aprajit Mahajan, David McKenzie and John Roberts (2011) "Does management matter? Evidence from India", World Bank Policy Research Working Paper no. 5573.

Boozer, Michael, Markus Goldstein and Tavneet Suri (2010) "Household Information: Implications for Poverty Measurement and Dynamics". Mimeo. World Bank.

Bruhn, Miriam (2011) "License to sell: The effect of business registration reform on entrepreneurial activity in Mexico", *Review of Economics and Statistics*, 93(1): 382-86.

Bruhn, Miriam and David McKenzie (2009) "In pursuit of balance: Randomization in practice in development field experiments", *American Economic Journal: Applied Economics* 1(4): 200-32.

Collins, Daryl, Jonathan Morduch, Stuart Rutherford and Orlanda Ruthven (2009) *Portfolios of the Poor: How the World's Poor Live on $2 a day*. Princeton University Press, Princeton, NJ.

Das, Jishnu, Stefan Dercon, James Habyarimana and Pramila Krishnan (2007) "Teacher shocks and student learning: Evidence from Zambia", *Journal of Human Resources* 42(4): 820-62.

De Mel, Suresh, David McKenzie and Christopher Woodruff (2008) "Returns to capital in microenterprises: Evidence from a field experiment", *Quarterly Journal of Economics* 123(4): 1329-72.

De Mel, Suresh, David McKenzie and Christopher Woodruff (2009) "Measuring microenterprise profits: Must we ask how the sausage is made?", *Journal of Development Economics* 88:19-31.

Drexler, Alejandro, Greg Fischer and Antoinette Schoar (2010) "Keeping it simple: Financial Literacy and Rule of Thumbs", mimeo. LSE.

Duflo, Esther, Rachel Glennerster, and Michael Kremer. (2008). "Using randomization in development economics research: A toolkit." In *Handbook of Development Economics, Vol. 4*, ed. T. Paul Schultz and John Strauss, 3895–3962. Amsterdam, NH: North Holland.

Fafchamps, Marcel, David McKenzie, Simon Quinn and Christopher Woodruff (2010) "Using PDA consistency checks to increase the precision of profits and sales measurements in panels", *Journal of Development Economics*, forthcoming.

Fafchamps, Marcel, David McKenzie, Simon Quinn and Christopher Woodruff (2011) "Female microenterprises and the flypaper effect: Evidence from a randomized experiment in Ghana", Mimeo. World Bank.

Frison, Lars and Stuart Pocock (1992) "Repeated measures in clinical trials analysis using mean summary statistics and its implications for design", *Statistics in Medicine* 11: 1685-1704.

Gibson, John, Jikun Huang and Scott Rozelle (2003) "Improving estimates of inequality and poverty from urban China's household income and expenditure survey", *Review of Income and Wealth* 49(1): 53-68.

Gibson, John and Bonggeun Kim (2010) "Non-classical measurement error in long-term retrospective recall surveys", *Oxford Bulletin of Economics and Statistics* 72(5): 687-95.

Gibson, John and David McKenzie (2010) "The development impact of New Zealand's RSE seasonal worker policy", World Bank Policy Research Working Paper no. 5488.

Glewwe, Paul, Michael Kremer, Sylvie Moulin, and Eric Zitzewitz (2004) "Retrospective vs prospective analyses of school inputs: the case of flip charts in Kenya", *Journal of Development Economics* 74: 251-68.

Imbens, Guido and Jeffrey Wooldridge (2009) "Recent developments in the econometrics of program evaluation", *Journal of Economic Literature* 47(1): 5-86.

Karlan, Dean, and Martin Valdivia (2011). "Teaching Entrepreneurship: Impact of Business Training on Microfinance Institutions and Clients". *Review of Economics and Statistics*, 93(2): 510-27.

Karlan, Dean and Jonathan Zinman (2011) "Expanding microenterprise credit access: Using randomized supply decisions to estimate the impacts in Manila", *Science* 332(6035): 1278-84.

McKenzie, David (2004) "Aggregate shocks and labor market responses: Evidence from Argentina's financial crisis", *Economic Development and Cultural Change* 52(4): 719-758

Miguel, Edward and Michael Kremer (2004) "Worms: Identifying impacts on health and education in the presence of treatment externalities", *Econometrica* 72(1): 159-217.

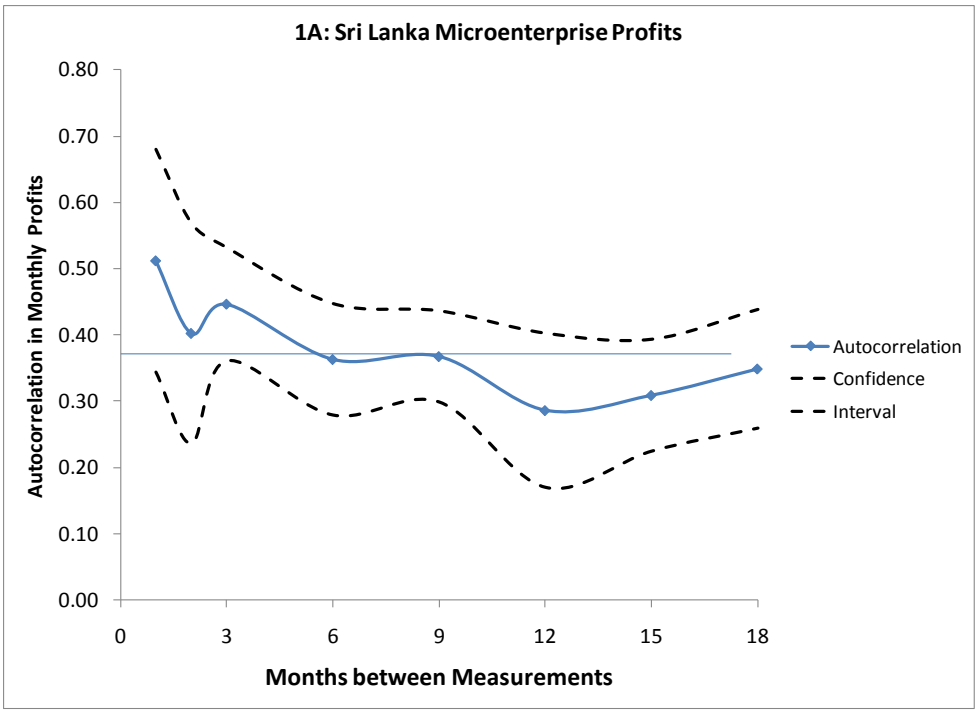
Samphantharak, Krislert and Robert Townsend (2009) “Households as corporate firms: Constructing financial statements from integrated household surveys”, *Econometric Society Monographs* (No. 46)

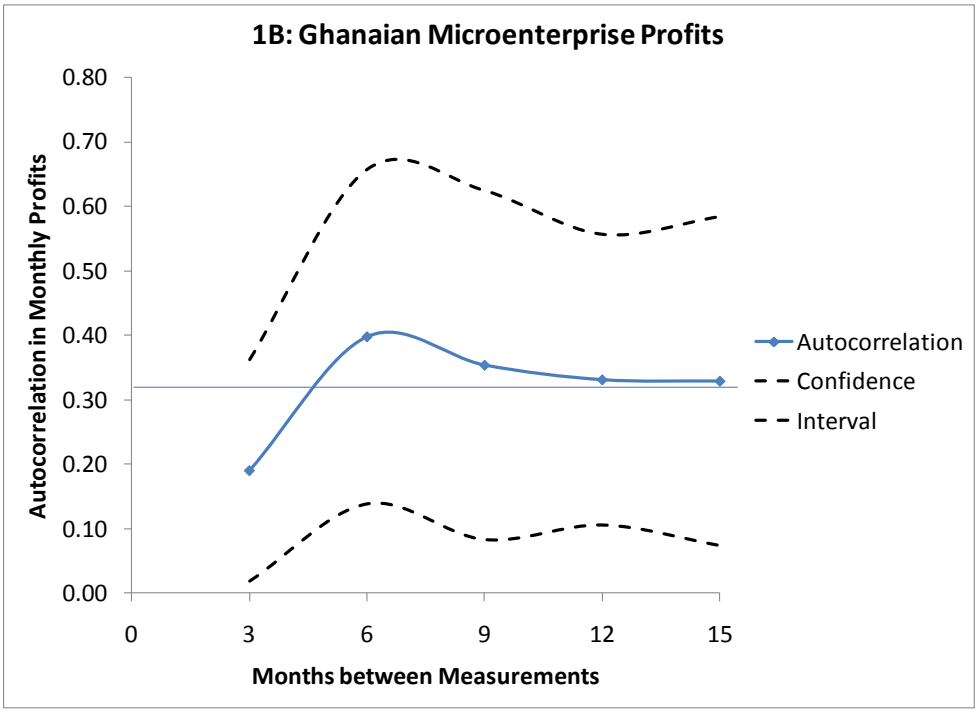
Vickers, Andrew (2003) “How many repeated measures in repeated measures designs? Statistical issues for comparative trials”, *BMC Medical Research Methodology* 3:22.

Woolcock, Michael (2009) “Toward a plurality of methods in project evaluation: a contextualised approach to understanding impact trajectories and efficacy”, *Journal of Development Effectiveness*,1(1),1-14.

Zwane, Alix Peterson, Jonathan Zinman, Eric Van Dusen, William Pariente, Clair Null, Edward Miguel, Michael Kremer, Dean Karlan, Richard Hornbeck, Xavier Gine, Esther Duflo, Florencia Devoto, Bruno Crepon and Abhijit Banerjee (2011) “Being surveyed can change later behavior and related parameter estimates”, *Proceedings of the National Academy of Sciences* 108(5): 1821-26.

Figure 1: Is the Autocorrelation approximately constant over different time horizons?





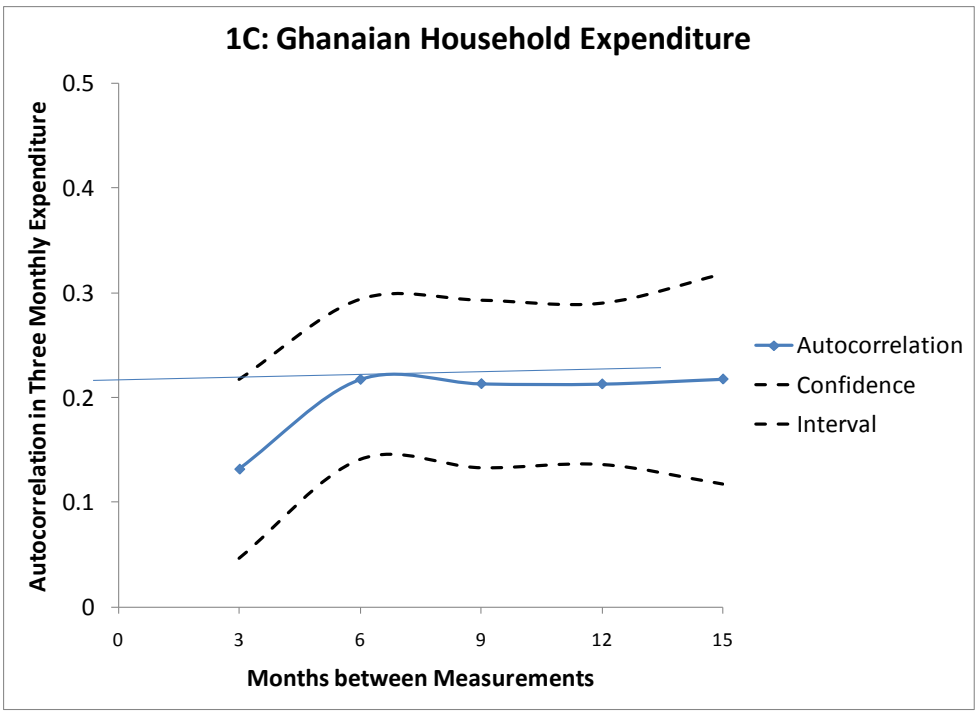


Table 1: Does the autocorrelation change with treatment status?

	Ghana Microenterprise Profits Control	Ghana Microenterprise Profits Treatment	95% C.I. for difference	Sri Lanka Microenterprise Profits Control	Sri Lanka Microenterprise Profits Treatment	95% C.I. for difference
Correlation between Baseline and:						
t=2	0.148	0.334	(-0.11, +0.46)	0.413	0.464	(-0.14, +0.25)
t=3	0.451	0.440	(-0.43,+0.33)	0.375	0.390	(-0.18, +0.13)
t=4	0.382	0.462	(-0.37,+0.44)	0.387	0.357	(-0.18, +0.11)
t=5	0.323	0.504	(-0.21,+0.47)	0.393	0.242	(-0.35, +0.07)
t=6	0.313	0.578	(-0.28, +0.64)	0.282	0.324	(-0.15, +0.20)
Average N:	325	344		220	318	

Notes:
Time periods correspond to calendar quarters.

Table 2: Examples of Autocorrelations for Other Economic Outcomes

Outcome	Source	Country	Time Interval	ρ
Household Income	Gibson and McKenzie (2010)	Tonga	6 months	0.38-0.47
		Vanuatu	6 months	0.19-0.21
Household Expenditure	Gibson and McKenzie (2010)	Tonga	6 months	0.12-0.33
		Vanuatu	6 months	0.35-0.53
	Boozer et al. (2010)	Ghana	6 months	0.32 (single-report)
				0.58-0.66 (separate reports)
	Gibson et al. (2003)	Urban China	2, 4 and 6 months	0.15-0.18
Individual Labor Income	EPH May and October 2002 (see McKenzie, 2004)	Argentina	6 months	Below median income: 0.25
				Above median income: 0.79
	ENE 2003:1-2004:1 (see Bruhn, 2011)	Mexico	3 months	Below median income: 0.29-0.31
				Above median income: 0.50-0.53
			6 months	Below median income: 0.22-0.31
				Above median income: 0.49
Math test scores	Das et al. (2007)	Zambia	1 year	0.68
	Andrabi et al. (2011)	Pakistan	1 year	0.61
	Banerjee et al. (2007)	India	2 years	0.59
Language test scores	Das et al. (2007)	Zambia	1 year	0.68
	Andrabi et al. (2011)	Pakistan	1 year	0.65-0.66
	Banerjee et al. (2007)	India	2 years	0.51

Table 3: How should sample sizes change if the autocorrelations differ in the treatment and control groups?

Panel A: Post Analysis

		Optimal ratio of treatment to control sample sizes		
ρ_{TREAT}	$\rho_{CONTROL}$	with 2 follow-ups	with 3 follow-ups	with 5 follow-ups
0.7	0.5	1.065	1.095	1.125
0.4	0.5	0.966	0.949	0.931
0.3	0.5	0.931	0.894	0.856
0.2	0.5	0.894	0.837	0.775
0.2	0.8	0.816	0.734	0.655

Panel B: Difference-in-Differences Analysis with one pre-treatment survey

		Optimal ratio of treatment to control sample sizes		
ρ_{TREAT}	$\rho_{CONTROL}$	with 2 follow-ups	with 3 follow-ups	with 5 follow-ups
0.7	0.5	0.775	0.775	0.775
0.4	0.5	1.095	1.095	1.095
0.3	0.5	1.183	1.183	1.183
0.2	0.5	1.265	1.265	1.265
0.2	0.8	2.000	2.000	2.000

Panel C: Ancova analysis with one pre-treatment survey

		Optimal ratio of treatment to control sample sizes		
ρ_{TREAT}	$\rho_{CONTROL}$	with 2 follow-ups	with 3 follow-ups	with 5 follow-ups
0.7	0.5	0.849	0.863	0.878
0.4	0.5	1.039	1.028	1.014
0.3	0.5	1.058	1.032	1.000
0.2	0.5	1.058	1.012	0.956
0.2	0.8	1.468	1.372	1.265

Table 4: How does the cross-sectional sample size required vary with correlation and rounds

Panel A: Sample Sizes required with Post Estimation

Pre	Post	p=0	p=0.25	p=0.5	p=0.7	p=0.95
0	1	1570	1570	1570	1570	1570
0	3	524	785	1047	1256	1518
0	5	314	628	942	1194	1507
1	1	1570	1570	1570	1570	1570
1	2	785	982	1178	1335	1531
1	3	524	785	1047	1256	1518
1	4	393	687	982	1217	1511
2	1	1570	1570	1570	1570	1570
2	2	785	982	1178	1335	1531
2	3	524	785	1047	1256	1518
3	2	785	982	1178	1335	1531
4	1	1570	1570	1570	1570	1570

Panel B: Sample Sizes required with Difference-in-Differences

Pre	Post	p=0	p=0.25	p=0.5	p=0.7	p=0.95
1	1	3140	2355	1570	942	157
1	2	2355	1766	1178	707	118
1	3	2094	1570	1047	628	105
1	4	1963	1472	982	589	99
2	1	2355	1766	1178	707	118
2	2	1570	1178	785	471	79
2	3	1309	982	655	393	66
3	2	1309	982	655	393	66
4	1	1963	1472	982	589	99

Panel C: Sample Sizes required with ANCOVA

Pre	Post	p=0	p=0.25	p=0.5	p=0.7	p=0.95
1	1	1570	1472	1178	801	154
1	2	785	883	785	566	114
1	3	524	687	655	487	101
1	4	393	589	589	448	95
2	1	1570	1413	1047	665	117
2	2	785	825	655	430	78
2	3	524	628	524	351	65
3	2	785	785	589	373	65
4	1	1570	1346	942	578	98

Note: Power calculations calculated for hypothetical experiment with control mean and standard deviation of 100, treatment effect size 10%, and size 0.05.

Table 5: How does power vary with n and T holding nT fixed?

Cross-sectional Sample n	Number of Pre-treatment rounds (m)	Number of post-treatment rounds (r)	p=0	p=0.25	p=0.50	p=0.75	p=0.90
Panel A: Post							
1000	0	1	0.609	0.609	0.609	0.609	0.609
500	0	2	0.609	0.516	0.447	0.394	0.368
250	0	4	0.609	0.394	0.293	0.237	0.213
Panel B: Difference-in-Differences							
500	1	1	0.201	0.252	0.353	0.609	0.942
250	2	2	0.201	0.252	0.353	0.609	0.942
250	1	3	0.162	0.201	0.278	0.491	0.865
100	5	5	0.201	0.252	0.353	0.609	0.942
Panel C: Ancova							
500	1	1	0.353	0.372	0.447	0.667	0.952
333	1	2	0.446	0.405	0.446	0.636	0.932
250	2	2	0.353	0.339	0.410	0.641	0.948
250	1	3	0.491	0.394	0.410	0.575	0.889
100	5	5	0.353	0.299	0.379	0.622	0.945
100	1	9	0.564	0.274	0.249	0.332	0.604