Structural Modelling in Policymaking

Orazio Attansio[1], Debbie Blair[2]

November 2018

[1] University College London and Institute for Fiscal Studies
[2] University of Oxford and Institute for Fiscal Studies

Suggested Citation: Attansio O, Blair D, 2018 Structural Modelling in Policymaking? CEDIL Inception Paper 9: London.

About CEDIL: The Centre of Excellence for Development Impact and Learning (CEDIL) is an academic consortium supported by UKAID through DFID. The mission of the centre is to develop and promote new impact evaluation methods in international development.

Corresponding Author: Orazio Attansio, email: o.attansio@ucl.ac.uk

Table of Contents

Abstract

Structural modelling, that is the use of behavioural models to add a framework to the decision problem of an agent, is a useful yet underused tool in evaluation. This paper provides a general introduction to structural modelling, as well as an overview of other commonly used evaluation techniques in Economics and other social sciences. It then goes on to show with three key case studies, how structural models can be used to enrich the findings from randomised control trials. The case studies cover a wide range of policy questions: examining demand for health products in Kenya, incentivising teachers to attend school in India, and evaluating conditional cash transfers for education in Mexico. The case studies show how structural models add to our understanding of the mechanisms behind a given treatment effect, how the findings may change when the policy is rolled out under different circumstances, as well as allowing for the evaluation of different policies that were not originally trialled. The common pitfalls of structural models are discussed, with guidance provided throughout on how to conduct sensitivity analysis and model validation. It is hoped that this paper will persuade other researchers to use structural models, in conjunction with randomised control trials, that will lead to improved evaluation results, a deeper understanding of important problems and better informed policymaking in the future.

Section 1

Introduction

In recent years, the quantitative evaluation of policy interventions has received renewed attention. From a policy perspective, assessing whether a given intervention is effective in achieving its stated goals is obviously important. From an academic point of view, impact evaluation may be difficult if the design, allocation and implementation of a given intervention is related to the problem it aims to solve. In general terms, establishing causality among observed correlations and solving the attribution problem can be very hard. For this reason, and because *ex ante* it is often impossible to establish the impact of a given intervention, much attention has been given to Randomised Control Trials (RCT) and other methods that, under certain conditions, allow us to estimate the impact of an intervention comparing a 'treatment' sample that receives the intervention to a 'control' sample that does not.

However, while the identification of the causal relations and solving the attribution problem rigorously are desirable objectives, estimating the impact of an intervention on certain outcomes of interest is not sufficient for the effective and efficient design of policies. Policymakers need to understand the mechanisms behind the impacts of a given intervention and the results of an evaluation. This paper will discuss ways in which behavioural models can be used to address the key issues of using and extrapolating the results of policy evaluations. Such an analysis, which is often referred to as 'structural modelling', is an important complement to policy evaluation and it is crucial for the extrapolation of a given set of results and to use such results for policy design. These issues are relevant for the treatment effects approach to policy evaluation used in many disciplines, including psychology, biostatistics, epidemiology, educational research and economics.

The promise of structural modelling has been recognised for a long time within Economics: the evaluation of policies under an explicit framework of human behaviour. Structural models make explicit assumptions about the preferences, technology, available information, constraints and the rules of interaction between agents. Such a structure then provides the necessary framework to enable: (1) a theory driven interpretation of the empirical evidence, giving insights into the mechanisms behind the observed impacts; (2) forecasting the effects of modifications to existing policies or implementation of existing policies in new contexts; (3) the combination of evidence from multiple studies. As it should be obvious, the validity of the results obtained within a structural analysis depend on the validity of the models used to achieve a set of results and the validity of the assumptions used to identify empirically the parameters of such models. A transparent use of structural model should lay out clearly the assumptions made, both in terms of the type of model used and in terms of the empirical variation which is used to identify and estimate the model. Such a use would also make clear that, in some cases, to use a structural model within an evaluation

context, it is necessary to specify parametrically the main parts of the model. In other situations, instead, a non-parametric or semi parametric specification is sufficient. By the same token, in some situations it is sufficient to identify and estimate only some component of the model, as it is the case in the sufficient statistic approach we discuss below.

Historically, progress in the use of structural models for policy evaluation has been hindered by the imposition of what has been viewed by many as implausibly strong identifying assumptions. The rise in the popularity of Randomised Controlled Trials (RCT) to identify treatment effects in social sciences can be seen as a response to this issue. Approaches based on treatment effects estimated by RCTs or natural experiments are appealing: they require fewer assumptions in relation to the assignment of the treatment, functional form, and the exclusion and distributional properties than do structural methods. However, they can only answer a narrow sub-set of the questions that are relevant for policy design and that are theoretically possible to answer using structural methods.

There are some key shortcomings of RCTs. Firstly, many important policies and interventions are not amenable to evaluation using natural experiments or conducting a randomized trial. Secondly, even among those that are amenable, scope for interpretation of the estimated effects, as well as their external validity and contribution to existing knowledge can be very limited.

A promising direction for methodological innovation and future research in the field of policy evaluation is combining the strengths of structural modelling and RCTs approaches. Over recent years, a number of seminal studies have demonstrated the potential of this approach to broaden the range of questions that can be answered within policy evaluation while maintaining the high standards for credible identification, which has been of central importance for the proponents of RCTs

Randomized trials can be used to validate the assumptions that are necessary in the estimation of structural models. The validated models can then predict responses to policy modifications, extrapolate findings to different situations, predict long-run effects and estimate impacts of policies not amenable to randomization (e.g. infrastructure projects). Randomised trials can also be used to estimate more realistic versions of the structural models used in the evaluation exercise.

Structural models can be used to make findings from randomized trials more informative. We can learn about the mechanisms driving the effects, and see what happens when we change a different aspect of the environment. For example, Attanasio et al. (2015) estimate a structural model of skill production to find that the positive impact of an early childhood stimulation programme, trialled using a randomized design, is explained by increases in the level of parental investment rather than by changes in the way in which skills are produced.

There are challenges to tailoring the specific structural parameters being estimated to the policy questions of interest and of finding opportunities for credible identification. The cost to building richer models of behaviour is that this then requires making increasingly strong assumptions. A profitable approach to avoid using too strong assumptions in the specification and estimation of structural model is the collection of

rich data. This strategy is particularly relevant when a given policy is evaluated through an RCT, which requires a data collection exercise. Data collected with an RCT should include not only outcomes but also a number of variables that could be used to estimate more realistic model. The structural model to be used should drive the data collection decisions.

Some policy questions may not require estimation of a fully specified model[1]; the key to effectively combining RCTs and structural methods in policy evaluation is striking the right balance between complexity and robustness of underlying assumptions to answer a wider set of questions than either method alone can, while maintaining a high standard for credible assumptions.

There are also examples where the simple use of an RCT hinders important results that can be uncovered using a structural model. An example of such a situation is in a recent paper by Attanasio and Pastorino (2018), who use the data from one of the most celebrated RCT in economics, the evaluation sample of Progresa, to which we will be referring to in Section 4.3, for a different question. When the Conditional Cash Transfer program PROGRESA was introduced in Mexico in 1997, some commentators and policy makers worried that the injection of this cash targeted to disadvantaged families could be partially expropriated by shop keepers who could increase retail prices in the villages that received the grant. The availability of a rich survey collected to evaluate with an RCT the intervention could be used to answer that question. The evaluation sample included 506 villages, 320 of which were randomly assigned to the program with the remaining 186 excluded from receiving the conditional cash transfers for a 2 years period and used as a control group in the evaluation. A number of papers, such as Hoddinott et al. (2000), compared prices for various commodities in treatment and control villages and found no significant differences. Attanasio and Pastorino (2018) notice that prices for basic commodities such as rice, kidney beans and sugar, are not linear and exhibit strong quantity discounts. They then construct a model of price discrimination that predicts quantity discounts and estimate it using the PROGRESA data. The model fits the data well; it also predicts an increase in quantity discounts. Attanasio and Pastorino (2018) then show that the tilting of the price schedule[2] is observed in the RCT data suggesting that the no effect on average prices actually implies an increase at low quantities and a decrease at high quantities. Such a price shift, which can be explained by the structural model, has important distributional consequences and yet is undetected by a narrow and simple use of an RCT. This an example of the use of a structural model, which uses experimental data and can, at the same time, point to aspect of the data that might be ignored and interpret them.

In summary the two main aims of the paper are to:

1. Provide an overview of structural modelling as a useful research tool for evaluation methods for:

[1] A fully specified behavioral model would include all the relevant agents, their preferences, beliefs and constraints; as well as fully modeling how they interact with each other and the surrounding environment.

[2] A price schedule shows the relationship between prices offered by sellers and how they vary by quantity. A tilt of the schedule results in a change in its gradient.

(i) understanding the mechanisms of change.
(ii) extrapolating policy evidence over the long run or to different circumstances including different interventions,

2. Show the potential of structural modelling methods to:
(i) Complement other methods of evaluation.
(ii) Widen the range of questions that can be studied within evaluation.

We do this by first discussing the goals of policy analysis to highlight how valuable structural models can be in fulfilling these aims. We will give an overview of different approaches used in economics, namely structural models, RCTs, instrumental variables and sufficient statistics. This analysis will be followed by a more technical exposition of structural models, as made clear with three case studies[3]. We then relate structural modelling to approaches used in other fields, specifically mediation analysis in psychology. The paper will then conclude with a discussion of the challenges of structural modelling and future steps that we believe will help further our understanding of behaviour and improve policy.

[3] These case studies are: *Short run subsidies and long run adoption of new health products: evidence from a field experiment*. (Dupas, 2014); *Incentives work: getting teachers to come to school* (Duflo, Hanna & Ryan, 2012); *Education Choices in Mexico : Using a structural model and a randomized experiment to evaluate Progresa* (Attanasio, Meghir & Santiago, 2011) .

Section 2

Why do we do Policy Analysis?

Proper analysis is crucial to understand the impact of any policy intervention and should therefore play an important role in policy design. Such analysis should *"evaluate the impact of implemented interventions on outcomes including their impacts on the well-being of the treated and society at large".* This could involve evaluating both historical and existing policies and examining their consequences, as well as predicting the impact of future potential policies. We may want to know if the policy has had a welfare improving effect. To do so we need to construct the hypothetical state that would have happened if the policy had not been in place. We also may want to know the channels through which change occurred to better understand how the policy worked.

Often, a particular policy occurs in a specific location or with a specific population and in order to generalise the results we forecast the impact on different populations and in different situations. The policy may have occurred at a small or localised level and we need to predict the effects of a wide scale roll out. Furthermore, proposed interventions may have no historical precedent and no feasible way to be trialled - or there may be too many options to trial each one. Finally, the environment in which the policy will occur, especially in the context of developing countries, may be too dangerous or difficult to run a RCT. Ethical reasons may prevent the full implementation of a particular design. Thus, we often need to know more than just the difference between treatment and control in a RCT. We need estimates of the potential effects of the various options: to inform their cost-benefit analysis and to make informed decisions about what policy to roll out.

We need a good understanding of the channels through which a particular policy operates. If we know that policy X led to outcome Y, but we have no idea why or how, then extrapolation to different populations, especially on a wide scale, may be severely flawed. Firstly, any difference before and after the policy, or between two different populations may not hold when the change is implemented at a wider scale. Even ignoring the difficult issues of implementing a given policy at scale with limited resources rather than in a very controlled environment, widespread roll out of a policy might have general equilibrium effects that were not present in a trial - causing prices or other aspects of the choice environment to change. Secondly, there may be long-run effects that cannot be estimated within a short-run trial. We need knowledge of the structure of the choice problem to make long run inferences in the short run. Thirdly, if we know the channel through which a policy operates, we can then work to target directly that specific channel in a way that may be more cost effective than rolling out a full policy with redundant parts.

We believe that the nature of policy evaluation supports the use of structural models and yet they still remain underutilised. In the next section we review the main approaches that can be used in policy analysis: structural modelling (SM), randomised control trials (RCTs), instrumental variables (IVs) marginal treatment effects (MTEs) and sufficient statistics (SS). We then go onto to discuss in detail the structural model

approach, and show how it can be used in conjunction with RCTs or other methods to offer a more complete understanding of behaviour.

Section 3

Overview of Modelling Approaches

In this section, we discuss common methods to evaluate policies used both in Economics and in other disciplines. This is by no means an exhaustive overview of all methods. Instead, we chose to focus on methods that are related to the structural modelling approach discussed in this paper. The details of each approach highlighted below may be overly technical for the requirements of some readers – and as such we provide a brief summary here.

Structural models, as used in Economics, attempt to establish causality, by writing down an explicit model of behaviour. A model uses a number of parameters believed to govern behaviour. By making assumptions about the data generating process, it might be possible to estimate these parameters empirically and therefore to say with a degree of confidence if the relationship is causal. Structural equation models, used in a variety of disciplines other than Economics, are similar, using explicit assumptions to infer causality – but have less of a focus on the exact behavioural parameters underlying the change.

At the other extreme, we have randomised control trials – which experimentally induces exogenous variation. This external imposition (provided a number of other experimental conditions are met) allows the researchers to assume that changes in outcomes between treatment and control groups are due to the intervention of interest. Instrumental variables are related to this technique, in that the treatment designation is essentially an instrument. Researchers are probably more familiar with instruments used in different contexts, but it is essentially the same underlying idea. Instruments should be correlated with the input of interest, but not affect the outcome in any way, other than via its link with this independent variable. Treatment assignment is the ideal instrument, but there may be a wide range of potential instruments. This will vary depending on the specifics of the intervention of interest.

In between these two points we have mediation analysis, common to psychology that does not involve a fully specified behavioural model but attempts to look at the mechanisms of change through which an intervention occurs. We include a somewhat technical discussion for interested parties, but in summary mediation analysis attempts to quantify the amount of the estimated effect size into the part due to the treatment directly, and the part due to the indirect effects via another important variable (known as the mediator). The mediator can be thought of the channel through which change is occurring.

We also briefly discuss the sufficient statistics approach, albeit this is rarely been used in Economics outside the tax and welfare literature. It attempts to estimate the policy relevant parameters, which are themselves functions of deeper behavioural parameters (such as the ones specified in a structural model). In doing so, it suggests that these are

"sufficient" for policy makers to know, as they govern the response to the policy of interest – and that further knowledge of the underlying parameters are redundant. This is, of course, very context specific and to be useful requires identification of a sufficient statistic for each question that may not be generalizable to other cases.

We hope that the following discussion of each method introduces potentially new methods to readers but also allows a more nuanced understanding of how they relate to each other, and in what settings they may be valuable.

3.1 Methods common in Economics

The evaluation of policy interventions in economics relies on econometrics. Econometrics is based *"upon the development of statistical methods for estimating economic relationships, testing economic theories and evaluating and implementing government and business policy"* (pg.1 Wooldridge, 2015). It fuses economic theory and statistical methods in order to analyse existing datasets and experimentally generated data. There are two broad approaches to econometrics covered in this paper – one that uses theoretical frameworks to put more structure on the data and one that avoids the additional assumptions needed to justify using models and relies on variation that is defined as unrelated to the outcomes of interests. Following much of the literature we define as 'structural modelling' the first approach and 'reduced form modelling' the second. We also provide a brief overview of other approaches, sufficient statistics and instrumental variables that are related to these two views.

Structural Models

Structural models are often thought of as complex mathematical models that describe the behaviour of the subjects of treatment or, more generally, economic agents as solver of complex optimization problems. While this description is appropriate in some cases, it is not accurate. A structural model is one that uses theory to study the behaviour of agents, including the recipients (or not) of a given treatment. In some cases, a complex structural model is needed to describe such behaviour, as individuals might be responding to a complex set of incentives and might be interacting in complicated ways. But that is not always the case, as we illustrate below. In some situations, very simple models can be very informative and identify crucial elements of individual decision process that drive both participation to a given programme and its reaction to it. These elements inform individual choices and, ultimately, affect the impact of an intervention. Their identification is key to understand what drives the observed impacts and, potentially, to modify and re-design the intervention.

In its simplest terms, a structural model consists of a problem, typically defined by a set of equations, that individuals solve. Individual behaviour does not need to be assumed to be rational or the functioning of markets and other institutions to be efficient. It is however, necessary to describe what determines individual behaviour and, potentially, how individuals interact. Models used typically include specifications for individuals' tastes and preferences, the information individuals have access to, the constraints and

the institutions faced by individuals in solving their problems, the technical relations that determine the outcomes of interest. The empirical approach to a structural model seeks to identify and estimate the key preference and technology parameters of a decision problem, parameters that are invariant to environmental changes. It is worth noting that we use the term agents/individuals throughout but can refer to higher levels of aggregation, at the group or institution level.

A model relates a set of observable variables *y*, that are determined inside the model (endogenous) to a set of observable 'explanatory' variables *x* that are determined outside the model. It also links these variables to a set of unobservable shocks – $\boldsymbol{\varepsilon}$. These shocks are assumed to be uncorrelated with the *x* variables (that is they are exogenous). Moreover, these shocks are often assumed to enter the model additively. In some cases, the $\boldsymbol{\varepsilon}$ variables are assumed to be observed by the individuals in the model but unobservable to the researcher, while in other cases they are ex-ante unobservable to everybody. Often, some assumptions about the dynamics of the problem are also necessary. For instance, one needs to assume something about the correlation structure for the $\boldsymbol{\varepsilon}$ and $\boldsymbol{x}$ variables. These assumptions might be relevant as current observed variables might be informative about the value that future variables will take.

Economic theory attempts to define as accurately as possible how these elements are related. However, there will inevitably be things that are either not measurable, measured with significant error or simply unknown to the econometrician. For a model to be useful, it does not have to be complete and true, instead it needs to be a sensible representation of the world and to have reasonable predictions. To estimate the model, one is often forced to make a number of assumptions about the functional form of the relationship, the joint distribution of the variables and the properties of the unknowns in the error term. Taken together, these equations and assumptions consist of an economic model that can be used to make predictions about behaviour. Such model is obviously rarely realistic nor does it represent a complete description of individual behaviour. However, a well-constructed model can isolate and understand specific drivers of behaviour and be useful in predicting certain aspects of individual choices and the impacts of policies.

To be explicit, a well-defined model consists of the following key elements (Browning, 2012):

1. Well-defined objectives for agents.
2. Well-defined constraints.
3. Statements about uncertainty, what agents know and the beliefs of such agents about the uncertain outcomes.
4. How constraints, preferences and beliefs vary across agents.
5. A model of interaction between agents.
6. Some idea of which of the above elements will be invariant to changes in the economic environment.

When building a model, there is an inherent tension between a model rich enough to explain behaviour and simultaneously simple enough that it is feasible to estimate.

When writing the model down, it is important to be explicit about the different channels and interactions amongst variables. This approach allows us to see clearly what should and shouldn't be in the estimated equations, ideally with every parameter having a clear theory-based interpretation. It also makes it clear what are the assumptions under which the model and its parameters are identified and the data needed to estimate them.

In practise, there are often many different models and assumptions that could rationalise the same empirical regularities. One consequence of structural modelling is that some of the assumptions made in some situations have implications for other observables and these observables have to be explained through the same model. This leads to testable implications of the model and can be used to check if the model is a good description of behaviour.

Building a structural model requires an understanding of individual's behaviour, economic theory, and the use of rigorous statistical techniques. Even when the theory suggests a clear relationship among some variables, this is rarely exactly reflected empirically, due to omitted factors and statistical noise. A good model will be explicit about possible omitted factors, measurement error and the necessary identifying assumptions required to estimate the model. The modelling process involves making many choices - decisions that reflect reality as accurately as possible and that are ideally guided by good economic theory. Sometimes, however, some assumptions are made to simplify estimation. The structure placed on complicated phenomena is likely to be incorrect in some ways (either due to unobservables not accounted for, incorrect assumptions or other design flaws) and as such this situation requires testing the sensitivity of the model to alternative assumptions.

Then the model needs to be estimated and the parameters of interest identified. The identification problem, in this case, involves determining the conditions under which one can be sure that an equation fitted to the data is actually the equation one wants. In a perfect world with infinite data, we know the value of the true statistic of interest. The question here is whether we can work out what we want to know and if so then our parameter is identified[4]. After identification has been established one usually proceeds

[4] What is meant by the term 'identified' can be tricky to pin down. In classic economic theory it implies '*that a model satisfies technical conditions insuring a unique global maximum for the statistical objective function.*' Within applied econometrics though the phrase 'how a parameter is identified' refers instead to the idea that Keane (2010) phrases: "*What are the key features of the data, or the key sources of (assumed) exogenous variation in the data, or the key a priori theoretical or statistical assumptions imposed in the estimation, that drive the quantitative values of the parameter estimates, and strongly influence the substantive conclusions drawn from the estimation exercise?"* We want to find a unique solution to whatever problem is at hand. In the case of structural models how do we solve for unique values of the parameters of the structural model from the values of parameters of the reduced form model. If we can find many values of parameters that fit then we have an over identified model, if we can't find any then it is under identified. Researchers should spend a lot of time thinking about whether their model is identifiable- a process often easier said than done. The classic example is when you observe an equilibrium price and quantity of a good. This has been determined by the intersection of supply and demand lines. However, given we only see the price and quantity over time we cannot estimate both functions with one linear regression line with these two variables. We would need another variable that shifts only one of the lines in order to pin down the slope of that line.

to estimation that is finding the estimates of the population distributions given the finite sample available.

Once the researcher has built and estimated a model of equations that they believe adequately describes the behaviour of interest - the model can then be solved to analyse situations that might be different from those observed. For instance, one may wish to study the effect of an intervention different from the one observed. Often this step involves simulating data from the model and generating predictions about economic behaviour. The simulated data can be used to check the predictions against what our intuition tells us or what we know from real life observations.

The structural approach allows the estimation unobserved economic or behavioural parameters that we could not infer from a comparison of means. Theoretical models can be used to generate counterfactuals and to perform policy simulations and predictions. The researcher uses the estimated parameters to determine what would happen if elements of the environment changed or if the policy had not occurred. To be convinced of the models' accuracy we need to ensure that the structural parameters of our model are invariant to external environmental changes. We can also use structural models to test competing theories of behaviour - does one model explain/predict observed behaviour better than the others?

This article will not go in depth into the theoretical details of identification or estimation (for interested readers, a good review paper is Heckman & Vylatcil, 2007). Instead it will focus on structural modelling generally and show how it can be used to improve policy analysis. At this point it is important to highlight that models should be validated by the researcher. By validation we mean that the model should be assessed as to whether it fulfils its intended purpose. The two main areas on which to assess a model is how well it fits historical patterns and moments; and how reasonable are its out of sample predictions. This involves some degree of subjectivity of course, but any validation process will and is not a reason to not engage in this important step in structural modelling. For an overview of some common methods of validation we direct the reader to Keane (2010).

Despite its benefits, structural modelling is not often used to inform policy analysis. We believe that this is in part due to some genuine criticisms of the approach, but also due to lack of familiarity with its techniques, in conjunction to the perceived difficulty of structural modelling.

Firstly, the modelling assumptions and techniques used can appear overly restrictive and unrealistic. These assumptions are made not only about the potential causal channels, but also about the functional forms of the key equations. The approach is sometimes seen as trying to fit a relatively simple and straightforward structure on something that is too complex to be modelled by a set of equations and assumptions. Secondly, the amount of maths and modelling that is required could be off putting for a non-academic audience (and even for some trained in Economics). Thirdly, the lack of computing power historically hindered the adoption of this technique, as it requires the ability to estimate a large number of parameters. Fourthly, many believed that once you

allowed for dynamics, selection effects and heterogeneity, it is no longer possible to identify and estimate deep structural parameters and this therefore makes the structural approach redundant. Some proponents of RCTs, in particular, might say that the presence of omitted variables, selection effects and simultaneity bias mean it is difficult to identify all primitive parameters in a meaningful way. When studying the effect of an intervention, therefore, the only alternative is to identify, with the help of exogenous variation controlled by the researcher, the impact of the intervention, even without fully understanding where such impacts (or lack thereof) come from.

Randomized Control Trials

RCTs are sometimes believed to be the gold standard in policy evaluation, offering an attractive tool for researchers to estimate the impact of interventions and establish causality cleanly. RCTs are believed to improve on a large number of alternative approaches, such as matching, quasi-experimental and instrumental variables, which are sometimes used to identify causal relationships in social sciences, often using observational datasets and relying on identifying assumptions that may or may not hold.[5] Advocates of this approach suggest that reduced form offers a simpler approach to econometrics than structural modelling. Here the econometrician does not build a comprehensive model of behaviour, at most they use theory to inform what variables are in `*y*' and in `*x*'. They then use the appropriate econometric technique to estimate the effect of *x* on *y*, without specific attention to the causal channel or other features of the decision making process. Often, X will be the treatment of interest – and will have been randomly assigned. This will ensure that assignment is independent of the outcome of interest and thus exogeneity is satisfied.

In the case of an RCT, evaluation involves comparing two or more groups (which ideally are on average identical in every attribute but the treatment of interest, although this does not necessarily hold). Differences in the distribution of outcomes between the two groups therefore identify the impact of the treatment. Often the analysis is focused on the difference between the two means, but in principle one can consider different moments of the distribution of outcomes.

For now, let us assume that the experiment was perfectly conducted, and focus on the difference between the means in the two groups. A statistically significant difference between the two means tells us that that it is an unlikely result if the policy had no impact. However, extrapolation beyond the mean difference and the interpretation of such a parameter is limited by the lack of an interpretative framework.

It will be hard to predict what will happen if we redo the policy in a different population, or at a much larger scale; nor do we know through what mechanism the policy operated or how the impact of the policy would change by changing some of the parameters of the policy. An RCT will answer a very specific policy question, but one that may not be

[5]For a review of methods currently used in applied econometrics to identify causality in observational datasets, see Athey & Imbens (2017).

widely applied beyond that circumstance. In order to test this the RCT would have to be repeated in different settings.

The treatment effect is an estimate that is not invariant to changes in the environment. To be valid we have to assume that the policy will not change any aspects of the environment - an assumption that is often difficult to defend. Many proponents of the reduced form approach believe the answers obtained with a RCT are more convincing as they require a smaller number of assumptions. There are many assumptions made in a structural model and in comparison an RCT may look attractive due to its simplicity. However, RCTs often have a lot of implicit assumptions. Michael Keane (2010) points out *"all econometric work relies on a priori assumptions"*, including randomised control trials and reduced form statistics. The link between the trial and the economic theory is not often explicitly stated in the RCT approach thus making it seem like they have fewer assumptions - whereas in reality there are many.

RCTs, on their own and not complemented by additional analysis and additional data beyond the outcome of interest, fail to meet the aims of policy analysis. They are limited to a narrow range of questions, typically evaluating one policy in one specific environment, they are not generally comparable across studies or to different populations/circumstances and they are not always strongly linked to economic theory. This obscurity means welfare comparisons and prediction to future changes are impossible.

Nevertheless, RCTs have grown enormously in popularity and have led to some positive developments - namely they have highlighted the need for specific attention to be given to the source of variation used in structural models and, more generally, to the identification of causal effects. This shift in attention has been a positive development. In fact, RCTs can be combined with structural models drawing on each of their individual strengths in a way that creates a stronger analysis. This will be highlighted in the case studies that appear later.

Instrumental Variables and (Marginal) Treatment Effects

One of the traditional approaches used in econometrics to address the issue of endogeneity of certain variables, such as the assignment of a treatment, has been the use of Instrumental Variable. The technique consists of identifying variation in a variable that moves the variable whose coefficient one wants to estimate while not affecting directly the outcome of interest. This variability can then be used to identify the effect that such a variable has on the outcome. In the present context, it will suffice to stress that the IV approach identifies the effect of an intervention by identifying one or more variables that determine the assignment of the program and have no direct effect on the outcomes of interest. IV uses the variability in treatment exposure driven by such variables to identify the impact of the treatment.

The idea of instrumental variables is an old one. Goldberg (1972), for instance, discusses early contributions by the geneticist Sewall Wright (1928) as one of the first examples of using variation in exogenous factors to identify the causal link between two variables.

Judea Pearl (2018) in his recent book discusses the example of the study of Jon Snow who in the nineteenth century was trying to identify the causes of cholera and was relating them to the quality of water. Access to high quality water is obviously related to many other variables linked to socioeconomic status and environmental quality which could give rise to spurious correlations that could be mistakenly interpreted as causal. Snow, who is often identified as the first epidemiologist, used the water company which was used by individual households to supply their water and the fact that there were two companies in London supplying water, both getting water from the Thames, but one upstream and one downstream from the sewage. Furthermore, both companies supplied most neighbourhoods, so that access to one company or the other was plausibly unrelated to other factors that could have affected cholera. Which water company was supplying the water of each household was therefore a valid instrument for water quality.

In this example and in many uses of Instrumental Variables, there is no explicit structural model of the variable that is being instrumented. However, the connection is implicit and the (untestable) assumption of the validity of a variable as an instrument, relies on a theoretical model. The assumption that is needed is that the instruments do not affect the outcome of interest directly but only have an effect on it through the effect they have on the 'endogenous' variable (or treatment). In some cases the validity of such an assumption is justified on an empirical basis and basic knowledge. The randomisation connected to an RCT is a good example of such a situation; the mechanism through which households would choose the water company supplying their water in London in the 1800s is another example. In other situations, the justification relies on specific behavioural assumptions. In both cases, however, the two features of a good instrument (the fact that it determines an endogenous variable and the fact that it can be excluded from the equation for the main outcome) are based on a model of the variables being instrumented. The other important feature we would like to stress is that such a model does not need to be specified and estimated fully and completely.

An important caveat to keep in mind when using instrumental variables, is that a mechanical use of the technique delivers the effect of interest when such impact is uniform in the population. However, in many situations, impacts might be heterogeneous. In such cases, one can think of different effects for different groups in the population. A recent approach in the academic literature has focused on *`marginal treatment effects'* (Heckman & Vytlacil, 2005), defined as the mean response of persons to treatment at a margin that is precisely defined to the question at hand and the method used. For instance, one can think that assignment to a treatment depends on its cost. The higher the cost of a treatment for an individual, the less likely it is that the individual will choose the treatment. Moreover, for a given cost, individuals with a larger impact will be more likely to choose the treatment.

The MTE approach goes beyond the simple comparison of means but does not involve a fully specified economic model. The aim is to *`unify the nonparametric literature on treatment effects with the econometric literature on structural estimation using a nonparametric analog of a policy invariant parameter'.* A marginal treatment effect estimates the effect of a policy but its value does not depend on the specific

circumstances that led to the treatment effect. It does not require a full specification of the underlying economic model - as is required by the structural approach.

Here we define the different treatment effects that are typically used in the RCT literature. The *Average Treatment Effect* (ATE) is the average effect over the entire distribution. The *Average Treatment Effect* on *the Treated* (ATT) averages over the distribution of impacts for those who are allocated to treatment. The ATU is the *Average Effect of the Treated on the Untreated,* roughly covering the changes in behaviour on those who were not treated directly but may still change their behaviours due to spill-overs effects. The ITT (*Intention to Treat)* estimator, the difference between those who were not intended to be treated (the control group) to those who were intended to be treated. The *Latent Average Treatment Effect* (LATE) averages over the distribution of impacts for those who switch into treatment because of a reform or after a change in value of some instrument affecting the decision to participate. This effect takes into account that treatment effects are heterogeneous within a population. Each of these is slightly different, operating over different margins. It can be difficult to definitively say which effect should be used in evaluation. That there are so many potential policy relevant treatment effects makes it hard to compare across studies or to know which is the most relevant.

In some cases it is clear which effect should be estimated and reported. If we want to evaluate a policy based on all of those who availed of a policy then we want the effect on the treated population. However, in some cases it is more relevant to consider the effect on all of those offered it - in which case the average effect might be more helpful. Policy makers may struggle to differentiate between these possible treatment effects and may make important decisions based on the wrong statistic - as a way to forward their own agenda, or simply by not understanding what the treatment effect is telling them. Heckman & Vytlacil (2005) show how all of these parameters can be written as a weighted average of the MTE, which then acts as a coherent framework for the differing effects.

The marginal treatment effect approach is the effect of a treatment on the marginal individual entering treatment - defined as the gain from treatment for individuals who are shifted into or out of the treatment by a marginal change in the cost of treatment (or the instrumental variable used in estimation). Heckman went on to extend this parameter by considering it to be the effect of the gain from treatment due to a marginal change in the propensity score (i.e. the predicted probability of treatment). Carneiro, Heckman & Vytlacil (2010) argue that policy changes on the margin are easier to identify and estimate than the other aforementioned treatment effects due to their limiting normal distribution, which has all the normal properties most researchers are comfortable using.

Sufficient Statistics

A related alternative is the sufficient statistic (SS) approach, as reviewed by Chetty (2009). The sufficient statistic literature proposes that many questions can be answered by estimating a smaller set of key parameters using programme evaluation models. Instead of estimating the deeper level parameters they combine various top-level

elasticities or functions of elasticities. Chetty (2009) argues that this approach *"combines the advantages of reduced-form strategies, transparent and credible identification - with an important advantage of structural models - the ability to make predictions about counterfactual outcomes and welfare".* It is proposed as a bridge between reduced form and structural modelling. The SS approach requires less data and variation to identify these marginal effects. In the SS framework, we do not need to identify all the model's primitive parameters, and thus it requires weaker assumptions than using other design based empirical methods. It can also be implemented in scenarios when we do not know the full positive model of behaviour, or when we are uncertain about the nature of this behaviour. The SS approach obviates the need for a full structural understanding of behaviour, requiring only a subset of structural parameters for predictions. However, some major drawbacks of this approach are that it often requires a new 'sufficient statistic' for each question at hand, it can be easily misapplied to situations and we still will lack full understanding of the underlying causal mechanism.

The approach has been most commonly used in public economics, in particular for taxation, social insurance and behavioural finance models. Chetty (2009) gives a number of examples of how high-level elasticities can be used to measure the welfare effects of policy changes. For instance, the marginal welfare gain from increasing income tax rate can be expressed as a function of the elasticity of taxable income even though taxable income is a function of complex decisions on wages, hours and training (Feldstein, 1999).

The approach reviewed by Chetty (2009) is related to that taken by Attanasio et al. (2018), who discuss the labour supply elasticity to wages, which could be informative in evaluating the effects of changes in taxation. Attanasio et al. (2018) present a structural life-cycle model of labour supply and show that certain elasticity concepts (such as the Hicksian and Marshallian static elasticities), can be evaluated with just some of the structural parameters of the model. These parameters can be empirically identified without estimating the whole model and using certain sources of variation in wages to pin down the relevant margins. Obviously, a complete answer to the question of the effect of a change in wages on labour supply needs the estimation of the whole model, which involves solving difficult identification problems. However, the static elasticities can provide bounds to the overall elasticities and the complete answer. What is important to stress is that some answers can be obtained with a limited number of assumptions.

One of the few applications of this approach outside the typical realm of public finance has been in a recent evaluation of Head Start (Kline & Walters, 2016). Head Start is an early childhood education program rolled out in the US in 1965. Today it involves a year-round program offering education, health and nutrition services to disadvantaged children and their families. This paper aimed to evaluate its cost effectiveness, taking into consideration substitution patterns. When a child is offered a preschool place, they may move from another form of public childcare, or from no childcare. If a child moves from one form of publically funded body to Head Start, the cost of implementation would not be as large as is normally included in cost-benefit analysis as they would have been using federal money anyway. In their analysis, they are able to derive sufficient statistics for the impacts and costs of Head Start (the policy relevant parameter) when

publicly subsidized alternatives are available. They find that those who would have attended preschool anyway are less likely to benefit from Head Start, whereas those who would have stayed at home accrue the largest gains.

The expression for the causal effect on child outcomes is a 'sufficient statistic' – a variant on the LATE concept when considering program expansions. This parameter is an average effect of Head Start participation relative to the next best alternative, regardless of whether that alternative is a competing program or home care. This local average treatment effect is identified by a randomized experiment with imperfect compliance when the experiment contains a representative sample of program compliers. These features of the experiment also mimic the true behavior of a population in this scenario, which adds to the results' validity.

When considering reforms that change selection into the program, the policy relevant parameter is a variant of the MTE concept. In this sense, even though a stylised model is built, to analyse the effects of changes in policy does not require full estimation of the underlying parameters, but rather of these higher-level effects.

3.2 Methods common in other fields

Economics is one field that deals with the evaluation of policy or more generally with the attribution problem. However, many other disciplines, such as psychology, sociology, epidemiology, deal with the same issues and have methods that are not wholly dissimilar to what economists call structural models. Here we provide a brief overview of two major approaches – mediation analysis and SEMs (structural equation models). The common threads across different fields are the problems of causality, its empirical identification and what conditions must hold for some conclusions to be credible.

Mediation Analysis

Mediation analysis is a technique for identifying causal effects that is commonly used in social sciences and medicine. It has some commonalities with structural models, and this section will show how they are related. We start with a simple description of mediation analysis. Let's say we know that a certain treatment impacts an outcome, but we don't know the causal path that generates such impacts. Mediation analysis helps figure out the mechanism behind how we get from cause to effect.

Consider a treatment D_i, outcome variable Y_i, and an intermediate post treatment variable M_i known as a mediator. The mediator is thought to change the effect of the treatment. Mediators have potential outcome $M_i(d$: the value the mediator takes when the treatment is d. The potential outcome is $Y_i(d,m)$: the value that the outcome takes when treatment is *d*, and mediator takes value *m*. The goal of mediation analysis is to identify the total treatment effect, and to separate it into the direct and indirect effect.

It should be apparent that this approach is reminiscent of the structural modelling approach – first the behavioural problem is considered, clearly defining the treatment, the confounders, the mediators and the potential pathways. Second, explicit attention is

paid to the assumptions that are sufficient for identifying a causal mechanism from the observed information. After we are confident that we can identify what we want to, we can estimate the different effects. It is good practise to conduct sensitivity analysis, to see to what extent the findings rely on key assumptions. This is particularly important to consider, especially with non-refutable assumptions, or in the cases when randomisation does not generate the required identifying assumption. We frame this discussion within the potential outcome framework. The total causal effect is the difference between one potential outcome that is realised when an individual takes part in a treatment and the other potential outcome that would be realised if they did not take part in the treatment:

$\tau_i = Y_i(1) - Y_i(0)$

The major problem is that this is an unobservable counterfactual - we don't actually see the alternative state of the world. Instead, we often focus on the average causal effect across the population in which some people receive the treatment and others do not.

$$\bar{\tau} = E(Y_i(1) - Y_i(0))$$

When the probability of receiving treatment is unrelated to the characteristics of the individual, then the average causal effect can be identified as the observed differences in mean outcomes between treatment and control. This is what is typically calculated in reduced form analysis.

We can extend this to include a mediator variable, $M_i(d)$ where d=1,0. The causal mediation effect is:

$$\delta_i(d) = Y_i\big(d, M_i(1)\big) - Y_i\big(d, M_i(0)\big)$$

This is also known as the indirect effect (INDE). We fix the treatment to *d,* and vary M_i by the value that it would take under treatment and control for unit *i.* If the treatment D_i, does not affect M_i so that $M_i(1) = M_i(0)$, then $\delta_i = 0$ (there is no indirect effect).

There is also a natural direct effect (NDE) of the treatment:

$$\eta_i(d) = Y_i\big(1, M_i(d)\big) - Y_i\big(0, M_i(d)\big)$$

Which is the effect of moving from the control to treatment while holding the mediator fixed at the value it would have under treatment status *d.*

The total causal effect is related to the NDE and the INDE, in that we can decompose the total effect of treatment into the sum of a direct and indirect effect[6].

---

[6] In reality one does not observe these states for each individual person but rather estimates using average effects, whereby half the people receive treatment and half the people receive the control. The average causal mediation effect (ACME), for d =(0,1)

$$\bar{\delta}(d) = E[\delta_i(d)] = E[Y_i\big(d, M_i(1)\big) - Y_i\big(d, M_i(0)\big)]$$

For the average direct effect (ADE) at the population level we have

$$\bar{\eta}(d) = E(Y_i\big(1, M_i(d)\big) - Y_i\big(0, M_i(d)\big)$$

The total average effect can be defined as:

$$\tau_i = \delta_i(d) + \eta_i(1-d) = INDE_i(d) + NDE_i(1-d)$$

The definitions of the INDE and NDE make the goal of a causal mediation clear: to take the total effect and decompose it into its constituent parts. The indirect component represents a posited explanation (via the mediator) as to why the treatment works, whereas the direct component represents all other possible explanations.

As these objects inherently involve counterfactuals we need to make a number of assumptions in order to identify the effects. To identify the total effect, we just require randomisation of treatment. Given that this is akin to the RCT effect as discussed earlier, it makes sense that this requires the least amount of assumptions. To go beyond this, and to identify the direct and indirect path we need more conditions on the data generating process to produce unbiased results.

Imai et al (2010) uses the sequential ignorability (SI) assumption to identify effects. This assumption has two parts.

Assumption 1 (SI): The treatment is independent of the potential outcomes and potential mediators, conditional on a set of covariates (pre-treatment confounders):

$$\{Y_i(d', m), M_i(d)\} \perp D_i | X_i = \text{x}$$

Assumption 2 (SI): The mediator is ignorable with respect to the outcome, conditional on the treatment:

$$Y_i(d', m) \perp M_i(d) | D_i = d, X_i = \text{x}$$

where $0 < P_r(T_i = T|X_i = x)$ and $0 < p(M_i = m|X_i = x, D_i = d)$ for *d= 0 ,1* and all $x \epsilon X, m \epsilon M$.

The first of these two assumptions can be satisfied through randomization. The second requires that any pre-treatment covariates that affect both the mediator and the outcome must be conditioned on. The main problem is that there may be unmeasured confounders that will cause this not to hold. Additionally we cannot condition on post-treatment confounders[7].

For both of the SI assumptions, we need the same conditioning set X_i. If the only aim is to identify the total effect, then pre-treatment confounders are of secondary importance. However, if we care about mediation analysis, then collection of pre-treatment covariates is of critical importance. If we can also collect baseline measures of outcome and mediator, then the plausibility of sequential ignorability is much more

---

$$\bar{\tau} = E(Y_i(1, M_i(1)) - Y_i(0, M_i(0)) = \bar{\eta}(1) + \bar{\delta}(0) = \bar{\eta}(0) + \bar{\delta}(1)$$

If we make the no interaction assumption (that direct and causal mediation effects do not vary as a function of treatment status) the ADE and ACME simply sum to the average (total) causal effect.

[7] Pearl (2014) suggests that these confounders do not necessarily need to be pre-treatment but they just have to be unaffected by the policy.

likely. Here the theoretical analysis can inform the design of the experiment and what data is needed.

It is worth noting that second assumption is generally not satisfied in an RCT and requires strong additional assumptions, which are somewhat similar to those used in a structural model approach. It is also interesting to note within a structural model framework, the SI assumptions take a more explicit and only slightly different role, being expressed as assumptions about behaviour and data variability.

One main difference between structural modelling and the causal mediation approach is that recent advances in the latter have promoted the importance of non-parametric forms of identification. This allows them to be applied in a wider variety of circumstances, when identification does not rely on functional form assumptions such as linearity. As stated previously, a causal quantity is identified under a certain set of assumptions if it can be estimated with an infinite amount of data. If the set of assumptions for identification does not involve any distributional or functional form assumptions, it is said that the identification is achieved non-parametrically.

After identification and estimation is achieved, it is important to test the sensitivity of the results to violations of the assumptions and to test the likelihood of our assumptions holding in the particular sample. Sometimes this relies on logical arguments rather than empirical tests, but again this depends on the question under examination and the data available to the researcher.

An example of where mediation analysis has been used in evaluation is with the Perry Preschool project (Heckman, Pinto & Savelyev (2013) and Keele, Tingley & Yamamoto (2015)). This was an early preschool intervention program aimed at disadvantaged African-American children during the mid-1960s in Michigan. It was designed to test the effectiveness of preschool classes on a wide range of cognitive and non-cognitive outcomes.

The Keele et al (2015) paper is interested in the effect of the program on high school graduation, with the mediator of interest being the impact on cognitive skills (IQ at age 6-8). Children were randomised into two years of specialised pre-school classes that lasted 2.5 hours for 5 days a week or else in the control group they were excluded from the specialized classes. Treated students were also visited by their teacher at home as a means to engage their parents. Admission was randomised and compliance was perfect, thus the first assumption of SI was satisfied. There are a number of pre-treatment covariates such as mother's education, mother's occupation, if the father was present, mother's age, father's unskilled work, density of people at home, child's sex and baseline cognitive skills that can be included.

For the second assumption of SI to be satisfied, there should be no unobserved pre-treatment covariates that affect both cognitive skill and high school completion. While there were many observed measures of pre-treatment variables that might affect both, there could be other confounds such as parental depression which may reduce IQ scores but also reduce the likelihood of finishing high school. It is hard to completely rule out any possible missing variables – which is one of the main reasons we need sensitivity analysis to check the robustness of the results. The authors calculate the

estimated ACME, ADE and the total effect. The ACME is the indirect effect, the portion of the average total effect that is transmitted through higher cognitive ability, and the direct effect is the remaining portion of the Perry Program effect attributable to all other possible causal channels. They find that the ACME for cognitive ability is approximately 0.069, whereas the average total effect is 0.224 and the ADE 0.169. These effects are dependent on the SI assumptions holding. To test the likelihood of the validity of the SI assumptions, the authors conducted a sensitivity analysis and concluded that it was likely that they did hold.

In comparison, the Heckman et al (2013) paper was not as explicit in discussing if the assumptions required for mediation analysis were valid. The strength of their paper lies in the range of possible mediators that they examined. They first examine the treatment effects on improvements in cognitive skills, externalising behaviour and academic motivation. They then estimate the relationship between these skills and later life outcomes and decompose these outcomes into components due to treatment induced changes in the different skills (or another unobserved "other factors"). The outcomes they examine include educational outcomes: the performance on the California Achievement Test (CAT), receiving special education at school and mental impairment; labour market outcomes: income and employment; health behaviours related to smoking or drug use; marriage duration and crime outcomes. Using mediation analysis, they uncovered that the effect of the intervention operated primarily via increases in externalising behaviour. This is in line with the findings from the Keele et al paper, which finds a limited effect via higher cognitive ability on cognitive outcomes of interest.

It could be helpful to promote the use of structural modelling by drawing on its similarities with mediation analysis – as researchers from many disciplines are already familiar with this framework and realise its utility in evaluating and understanding problems. The Heckman et al. paper clearly shows the utility of decomposing treatments effects into its possible channels of change, to better understand the policy intervention. However, it is also useful to be explicit about the assumptions on behaviour and data availability that are necessary to identify the mechanisms posited by mediation analysis.

How would a structural model have approached the Perry Preschool Project? Perhaps it could have gone further to examine through what channels the intervention affected the factors identified by Heckman et al as possible mediators. Policy makers may want to know whether it was the increased classroom time via the preschool classes, or whether it was the teacher home visit that mattered. It would be useful to examine which of these caused the desired change in outcomes. From a monetary and time cost perspective, going to each individual's house is probably not feasible at large scale, whereas having preschool classes for everyone is more scalable. However, if all the effects are driven through the home visit, then increasing preschool classes for all will not have the desired outcome, and the finding from the RCT will not be replicated.

It would also be interesting to look at even more intermediary measures of change. In the case of cognitive ability, yes, test scores changed but why? Was it via the increased instruction time, increased quality of instruction, changes in the beliefs of the

parents/child about what was expected/possible? A more detailed model of behaviour allows us to write down the number of different pathways and to really think through why the change is occurring. It may also help consider how to scale up the study – in fact the Perry Preschool Project involved only 128 students and has been used to inform the nationwide Head Start program. However, this nationwide program has not had the same success as Perry Preschool Project, suggesting difficulties with large scale roll out. It would be helpful to know why it has been difficult to replicate at scale. Structural modelling may be able to play a role in this.

Structural Equation Models (SEMs)

SEMs are often used in social sciences, with one of the earliest precursors being the work of Sewall Wright – one of the founders of the path analysis approach (Wright, 1921). His 1921 paper focused on guinea pig birth weight, but subsequently applied it to a wide range of agricultural phenomena. A short digression on path analysis is warranted here. Nowadays path analysis is conceptualised as a special case of SEMs, typically assuming that all variables are measured without error. Path analysis attempts to write down a graphical representation of the causal relations between variables, which corresponds to a system of equations. In this sense, it is a flow diagram representing a structural model of behaviour. These flow diagrams can be very complex, representing numerous pathways and both unidirectional and bidirectional relationships between multiple observables and unobservable influences. Ultimately, the goal is to try first to represent graphically the system through which a causal effect could be occurring, and then using this explicit diagram to inform the analysis if the coefficients of interest can be identified.

Path analysis is also related to directed acyclic graphs (DAGs). DAGs originated in computer science, but have now become commonplace in epidemiology and psychology. A DAG shows symbolically the dependence structure of multiple variables, and help researchers to think through the causal pathways they are attempting to identify. DAGs can identify conditional independence relationships implied by recursive product decompositions (Pearl, 1995). Essentially this joint probability distribution can be decomposed as a product of the conditional distribution of each node, given its parent nodes in the graph. This allows identification of the most likely directional link. Ideally, the DAG should be explicitly grounded in the conceptual understanding of the treatment of interest, with any key variables shown in the diagram, even if these variables cannot be measured. The graph is directed with arrows representing causal relationships, ones that are assumed to be structurally stable – so that intervening does not cause the association to change. DAGs do not have to be linear, nor do they have to be parametric. DAGs however, avoid simultaneity (i.e. there are no directed cycles), which distinguishes them from SEMs.

Both DAGs and path analysis are related to the structural approach. They require the researchers to formulate explicitly the pathways and the relationships between variables. SEMs also share this feature. SEMs do not have to be linear, nor do they have to assume homogenous effects but the most common method assumes linear SEM with constant coefficients. Typically, a SEM consists of two parts; the first being a set of

equations that describe the causal relations between the variables of interest. These variables are called latent variables when they are not directly measurable – many variables we are interested in are not directly measurable but instead we rely on proxy measures/indicators. The second part of a SEM is a measurement system that explicitly describes the relationship between the latent variables and the observable indicators that provide information on them.

The first part can be described by the following set of equations:

$$\eta_i = \alpha_\eta + B\eta_i + T\rho_i + \zeta_i$$

where η is the endogenous latent variable vector, ρ is the vector of latent exogenous variables, α_η a vector of intercepts and ζ_i a vector of disturbances or shocks to the system. Typically, we assume that the mean of the disturbance vector is zero and that the errors are not correlated with the latent exogenous variable. We wrote this system as being linear, but non-linear systems can be used as well. Moreover, we are abstracting from any temporal or dynamic effects, which can also be incorporated into the model.

This measurement part of the model underlies what psychologists call confirmatory factor analysis (CFA) – a means to link observed variables to their underlying unobserved latent variable. CFA is used frequently in a number of situations, often without understanding of its relationship to SEMs. The measurement system can be described by the following equations:

$$y_i = \alpha_y + \Lambda_y\eta_i + \varepsilon_i$$

$$x_i = \alpha_x + \Lambda_x\rho_i + \delta_i$$

where y_i is the vector of indicators for η_i, and x_i is a vector of indicators for ρ_i, with the other variables having their usual interpretation as intercepts, vector of coefficients of effects and error terms. The model typically assumes that means of the disturbances are zero, uncorrelated with each other and with the latent exogenous variables. It is often assumed that the matrices of 'loading factors' Λ_y and Λ_x are such that each indicator corresponds to only one latent factor, in which case the measurement system is called a dedicated system. The ε_i and δ_i can be interpreted as 'measurement errors' that prevent the observation of the latent factors.

The structural part of SEMs refers to the fact that the researcher incorporates causal assumptions as part of their model. These assumptions can either be "strong" such that a parameter takes a specific value (such as assuming zero when there is in fact no causal relation) or "weak" in that there is assumed to be a range of possible values. Depending on the nature of the assumptions made, the model is more or less restrictive and demands more of the data. This approach is very similar to the nature of the assumptions made in what economists call structural models. Imposing different assumptions leads to different models of causality, and ideally any assumptions made should be derived from theoretical models, prior studies, or another justifying source.

By making assumptions, researchers want the resulting model to have testable implications and for the structural coefficients to be estimable from the data (i.e. identifiable). Again, this resembles closely the structural approach taken in Economics.

The process of writing down SEMs, both in diagrammatic and equation format, making explicit what assumptions are being made, and how they aid identification, is at the heart of structural modelling being proposed in this paper. Structural models in Economics, as will be seen in the later case studies add more structure in their modelling than SEMs. In particular, the assumptions imposed on a model are typically derived from economic theory or, more precisely, translate assumptions about the type of behaviour of individual agents or their interaction. Having said that, structural models in Economics are not completely different from SEM in their goal of identifying causal pathways and deriving estimable structural parameters that are invariant to the treatment of interest. Both are useful in deriving estimates that can guide policy analysis.

Another resemblance between SEMs and structural models is that SEMs can be used with RCTs. Although in practice this does not seem to occur very often, Bollen & Pearl (2013) suggest that SEMs are a useful tool to *"help to determine (1) if the randomized stimulus actually affects the intended variable ("manipulation check"), (2) if the output measure is good enough to detect an effect, (3) if the hypothesized mediating variables serve as the mechanism between the stimulus and effect, and (4) if other mechanisms, possibly confounding ones, link the stimulus and effect."*

SEMs have rarely been incorporated in Economics for a number of reasons. They are often scapegoated, due to their misuse in other fields. Researchers oftentimes are non-explicit about the identifying assumptions they are making, and fail to conduct rigorous tests of any assumptions. However, good research using SEMs involves specification of the model, identification of the key parameters, estimation, testing both the model and its hypotheses, evaluating model fit and then using the model to make predictions. This is not so different from the structural approach in Economics, albeit the language and the type of model built are very different as will be seen in the case studies below.

Section 4

A Formal Structural Model

A fully specified structural model is hard to define in a general sense, as what a structural model is depends on the question under investigation. As mentioned already, a structural model is not necessarily complex. In some cases, simple relations can play that role. Moreover, even within a full model, there are still going to be missing variables, either unintentionally left out by the researcher or recognised to be important but not measurable. There are always simplifying assumptions made: a particular choice variable might be completely omitted, or one might condition certain choices on others. By '*fully* specified' we mean that the model's variables of interest are obtained as complete solution to individual optimization problems as a function of their current information set and, possibly, by the interaction of different individuals. The set of behaviours that can be represented in such a fashion is less restrictive than it seems at first sight. One can allow for rich preferences and limited information sets. One can also consider frictions and imperfections in the interactions among the agents. It has to be stressed that a structural model does not imply that the agents involved are fully rational or that they act optimally or that markets are imperfect. The structural model, however, can provide a coherent framework to represent the behaviour of different agents and their interactions.

As a reminder, in Table 1 we list the main components of a structural model that are likely to by included in most analysis. This list is by no means exhaustive, nor are all the components necessary to form an appropriate structural model. It is instead intended to give the reader an idea of what is required when coming up with a structural model of behaviour.

In the process of thinking about a model, one also needs to be thinking about what not to model. The list of omissions will always be long, but necessary. Individual decisions can be very complex and might need extremely rich structures to be represented in a coherent way which is consistent with observed facts. It is important to think about the model in terms of making it adequately complex in the correct dimensions and transparent enough to follow.

We provide three examples that should make clear how these different elements vary depending on the type of behaviour that is being examined.

Table 1: Common components of a model

Agents	Who are the key decision makers in the model, how many are there and how do they interact? Are we modelling individuals, groups or larger institutions?
Time Periods	The timing of the behavioural problem needs to be explicitly laid out. If it is a multi-period problem, the decision variables at each time period need to be stated, as well as clear exposition as to how variables evolve over time.
Preferences	The agents of the model will have a utility function. The components of this function should be explicitly laid out, as well as what the choice variables are. This utility function may be given a specific functional form, or may be given as a general case.
Constraints	The budget constraint needs to explicitly state. Most frequently thought of in terms of money – but it can be in the form of time, information or any other variable under which the individual must make a decision.
Information	What can the agents know (i.e. what is in their information set), and do they all know what the other agents know. How does information update, is it perfect,
Beliefs	What do the agents believe about the world? How do they change their beliefs in response to new information? How are their beliefs/expectations formed?
Uncertainty	What are the sources of uncertainty in the model and how do these sources impact choices and beliefs.
Broader Economic Environment	What are the external factors that might alter the behaviour of the economic agents? What are the external markets, institutions, and other parties (such as governments) that impact?

The three cases studies we discuss are from the economics literature and focus on social interventions in developing countries. Each of these examples differ in terms of

the scope and scale of the structural model. We cover each study in detail, including both a general overview of the study design, and the more technical elements of the model and identification of the parameters. This is to appeal to the wide range of parties that may be interested in understanding structural models, but who vary in their level of experience with quantitative methods.

The first case study looks at demand for bed nets in the short and long run, depending on the level of subsidy provided to the consumer. This case study is included to show how a structural model can be quite simple – requiring only one key demand equation from which individuals make purchasing decisions. We want to emphasise that structural models are not inherently complex or difficult. They can encompass a relatively simple structure that still generates a rich understanding of behaviour.

The second case study examined the effect of monitoring on teacher absenteeism. We include this specific example to show how the identification of the behavioural parameters enables going beyond evaluating the very specific intervention (in this case the payment structure of teachers), in order to predict a policy that might generate the same results, but at a lower cost. In a world where it is not feasible to run the many different potential interventions, the ability to extrapolate using unchanging structural parameters that have been identified is invaluable for policy makers.

The final case study is perhaps the most complicated in terms of the behavioural model, but is an example of the utility of structural models in determining the mechanisms of change and providing a rich set of findings. This model focuses on the education subsidies provided to households as part of the well-known PROGRESA initiative in Mexico, and shows how variation induced by an RCT can be used to help identify the structural model. This highlights our perspective that the two can be used as complementary methods of enquiry.

4.1 Case Study 1: Short Run Subsidies and Long Run Adoption of New Health Products: Evidence from a Field Experiment (Dupas, 2013)

A common policy tool in developing countries is to provide short-run subsidies for health products. By decreasing the price of a health positive product, practitioners hope to increase its uptake, benefitting the individual and potentially, through externalities, the community. There are a number of arguments in policy circles, for and against short-run subsidies, due to the effects on short-run usage and waste, the potential generation of expectations of receiving other products at subsidised prices, as well as long-run demand for the health products.

Subsidising a product could conceivably increase or decrease demand. If short-term subsidies increase demand and usage, and the product provides a positive health benefit, then there could be a learning effect through which people increase their value of the product, increasing the willingness to pay for the product in the long run. There may be spillover effects through learning on those who have not received the product, through network effects. The short-term subsidy may decrease usage if people do not

value (and are less likely to use) something they receive at a low price. Likewise overall usage may go down if people, who would not purchase the product because they do not need it, now receive it only due to its lower or free price. If usage decreases, then the potential learning effect will also be smaller.

Another possible negative side effect is that offering discounted prices for a product might generate an expectation of how much the product is worth. If the person takes the reduced price as a reference point then they may be less willing to pay for it in the future, thus decreasing demand.

It is clear that there are many ways in which a subsidy could potentially impact demand, both in the short and long run. It is important to consider what mechanisms are operating behind this change in demand. This case study will show how structural modelling can be a simple method for understanding the mechanisms behind a demand problem, as well for generating predictions about other policies that were not originally tested.

The normal method of evaluation would be to run an RCT, to test whether subsidies in the short run would increase or decrease demand in the long run. This particular RCT focused on the use of an Olyset long-lasting-insecticide-treated bed net (LLIN), a top of the range bed net assumed to decrease malaria.

In the first part of the experiment, individuals from six treatment villages received a voucher with a different randomly assigned subsidy, that resulted in bed nets prices ranging from 0$ to 3.8$. Participants had three months to purchase at a particular vendor, who could only sell Olyset nets to people with vouchers. This meant the only way to get this particular type of net (which was new and noticeably different to any older versions) was via the experiment. This was to prevent any spill over effects occurring via any other paths, other than the experimental imposed channel.

One year later, in the second phase, individuals were given the opportunity to purchase a second net at a fixed price. As listed previously there are a number of factors influencing this purchasing decision. There are learning effects, health benefits, and reference points which are generated by the cost of the first net; and negative effects due to decreasing marginal returns of the nets, given a household may already own one net.

The outcomes of interest at phase one were the demand at each price (as measured by those purchasing the nets using the vouchers) and the usage of the net (as measured by self-reported and observable measures two months after purchase). The outcome of interest in phase 2 was the number of people who purchased the net.

The reduced form results revealed that in the short run, there is quite a steep demand curve for LLINs, with sharp decreases in demand as the price increased, dropping from 97% for free nets, to 30% if the price rose above $1.50. Usage of the nets at both the two months and one year follow up did not seem to vary over the price paid.

This is evidence against the commonly cited claim that people are more likely to waste a product when they don't pay for it. They also found that people who were given a high subsidy voucher were more likely to purchase a net sooner. Those individuals who were

surrounded by a bigger share of people receiving high subsidies were also more likely to purchase their net, suggesting positive external effects to high subsidies.

In the long run, the authors examined if those who had received a high subsidy were more or less likely to buy at a higher price. Overall, there was evidence that there was a modest, yet insignificant positive effect of a high subsidy on phase 2 uptake. This suggests that the willingness to pay in high subsidy group increased despite the fact that they were buying a second net, with a lower marginal utility that would otherwise decrease the willingness to pay. The reduced form results also suggested a negative spillover effect in the long run. This could be that, if the nets reduced overall levels of malaria transmission, then the private benefit to purchasing a net decrease making it less attractive to buy a net.

The reduced form results reveal interesting patterns but do not reveal the breakdown of this overall effect on demand into its various mechanisms. This is a clear example of when building a structural model can help separate out the different mechanisms driving demand. Here there are potential positive learning effects and potential negative effects of the reference point. The results also suggest that exposure to highly subsidized neighbours reduced long-term adoption, which is another channel through which subsidies could reduce willingness to purchase in the longer run.

We will now give a summary of the model that Dupas builds and show how the RCT design enables identification of the key parameters. This is an additional benefit of combining RCTs and structural models. Without the RCT it would be difficult, if not impossible, to estimate the different parameters. We can also use the structural model to think about what might happen in the case of other health products. This allows us to extrapolate beyond this particular example and make informed predictions about other useful policies.

The demand equation consists of two periods and has one preventative health product. In each period the household will invest in the health product if the marginal benefit is greater than the marginal cost. Utility consists of two components: the intrinsic utility of the product and the gain-loss utility. Intrinsic utility is a function of outcomes, expected private benefits (depending on the disease burden, own vulnerability and beliefs about the quality of the product) and private costs. Gain-loss utility captures the reference-dependence nature of preferences. If households pay more than they expect for a product they will experience a utility loss. This simple model also has two forms of spill overs – health spillovers (when the disease risk is reduced) and information spill overs (signal of quality). Households make decisions based on the short run, which do not typically take into account the long run effects of their current choices.

In the first period, households do not know the quality of the product, they have no reference point for the price, and both the disease incidence and their vulnerability to the disease is high. Households can acquire the product at a price P, and will receive some information about its quality. They then are able to update their beliefs about the quality. In the second period, households are offered the products at a new price, taking the first period price as their reference point. The disease risk depends on adoption in period 1. Households are also less vulnerable to risk once they own one unit of the

product. They also know their own private health outcome, which acts as a signal for the quality of the product.

Households will invest in period 1 if the expected health gain is greater than a household-time specific shock, the price of the net, *a* which is the marginal utility of income and a function of the gain-loss utility.

$$E(R_t) > \varepsilon_{ht} + ap_t + f(p_{rt} - p_t)$$

The gain loss utility is assumed to be linear, taking the following form:

$$f(p_{rt} - p_t) = r\,.(p - p_r) \text{ if} > p_r\,,$$
$$f(p_{rt} - p_t) = r\,\lambda.\ (p - p_r) \text{ if } p < p_r,$$

where p_r is the reference price, and λ is the loss aversion parameter.

They also assume that $E(R_t) = E(m)\,\vartheta_t v_{ht}$, where *m* is the quality of the product, ϑ_t is the disease risk in the area and v_{ht} is the vulnerability of household to this disease risk at time *t*. If a household does not have a bed net the return to a new net is higher than if they already own one, capturing decreasing marginal utility.

At the start of period households have a prior belief about the quality *m,* that is drawn from a distribution $N(\mu, s_0^2)$. There is no gain loss utility in this period as the reference price is the price they are offered. With a disease burden of ϑ_1, and vulnerability v_t = 1, the household will buy the product if:

$$\mu\vartheta_1 > \varepsilon_{h1} + ap_1$$

If the product is purchased, individuals gain information about its quality, and they update their belief so that it comes from the distribution $N(\mu + l_1, s_1^2)$. Households who do not buy the product can also learn from other households who do have the product. The authors consider a learning process that depends on a network, a household with *n* owner households within a 250-meter radius. There are *N* households total in the radius, and *c* social contacts. The household has a chance of $1 - (1 - \frac{n}{N})^c$ to update to $N(\mu + l_1, s_1^2)$. If they do learn they update their purchasing decision and buy in period 1 if:

$$(\mu + l_1)\,\vartheta_1 > \varepsilon_{h1} + ap_1$$

In the second period, the disease risk will have changed based on the adoption in period 1. Specially $\vartheta_2 = \alpha\vartheta_1$, with $\alpha < 1$ if the local take up is above a particular threshold *t*. Now, households have updated beliefs about quality. Those who adopted in period 1

received a private signal $r_{1h} = m + v$ where $v \sim N(0, k^2)$ and update their belief using Bayes' law. They adopt in period 2 if

$$(\mu + l_1 + l_2)\vartheta_2 v_2 > \varepsilon_{h2} + ap_2 + f(p_1 - p_2)$$

where $l_2 = \frac{s_1^2}{\sqrt{k^2+s_a^2}} . \frac{r_{1h}-\mu-l_1}{\sqrt{k^2+s_a^2}}$: that is how much they update their belief based on mean quality and $v_2 < 1$ indicate that the returns to the second Olyset net are lower than the first.

Households that did not adopt the product in the first period, will adopt in the second period if:

$(\mu + l_3)\vartheta_2 > \varepsilon_{h2} + ap_2 + f(p_1 - p_2)$

where $v_2 = 1$ and $l_3 = 0$ for those who did not learn about the product, and $l_3 = l_1$ for those who did learn from their neighbours.

These simple demand equations govern the decision making of households in the short and long run.

The authors use maximum likelihood estimation to estimate the model, assuming that the distribution of preference shocks is logistic. They allow for village fixed effects and estimate the five key parameters.

- μ : the prior belief on quality of nets.
- l_1: how much the mean prior on quality changes upon learning from product characteristics, which is also transmitted via social learning.
- l_2: how much the mean prior on quality changes upon experimentation over a year.
- r : the reference dependent parameter.
- a : the marginal utility of income.

In order to identify these parameters the exogenous variation induced in the experiment is used. They also need to make a number of assumptions about the other free parameters in the model. It is common practise to fix the level of some parameters based on knowledge from other studies. [8]

[8] The parameter value assumptions made in this case were as follows:

- The threshold for health spill overs is fixed at t=0.6, with a strength of $\alpha = 0.8$. That is if net coverage reaches at least 60% amongst our households, the disease risk after a year is reduced by 20%. These parameters were taken from the medical literature on this area of research, such as Hawley et al. (2003) and Killeen et al. (2007
- The return to the second net is 85% of the first one, that is $v_2 = 0.85$, and $v_1 = 1$.
- The loss aversion parameter, λ = 1.6, based on the findings of a recent review of this literature.

The random variation induced by price variation (which generates both random variation in both experimentation price and reference points), along with the random variation in the share of neighbours within a 500m radius receiving the high subsidy, enable separate identification of learning from experimentation l_2 and reference dependence r.
This random variation in the share of neighbours within the 250 m radius receiving a high subsidy (and thus those who immediately purchase the product upon receiving the voucher) enables estimation of l_1 from phase 1 adoption design. The random variation also affects ownership at start of phase 2 but this alone cannot be used to estimate l_1, from phase 2 adoption of those who did not purchase in phase 1, since by then it is confounded by health spill-over (i.e. a change in disease burden).

The authors find that the reference dependence parameter, r, is small economically – totalling 6.6% of the estimated value of the marginal utility of income. In contrast, the total learning effect after a year of experiment $l = l_1 + l_2$ is quite large - corresponding to an increase in perceived quality of 41%. This increases demand by the equivalent of a price decrease of 30%. The short-term learning from product characteristics is about 31% of the total learning effect; this is the aspect of learning that can be transmitted via network effects. The learning about the product from neighbours increases demand by about as much as a 9% price drop.

One of the biggest criticisms of the structural model approach is that in order to derive estimates such as the ones above, a large number of assumptions are required. It is necessary and important thus to test the sensitivity of the results to the assumptions made, in order to have confidence in their legitimacy. In this example, the results were robust to different levels of the parameters whose levels had been fixed.[9] The authors were also able to derive which assumptions mattered for identification of different parameters. For example, the estimate of year-long learning (l_2) depends on diminishing returns assumptions.

These results show how a relatively simple structural model helps to decompose an effect into its underlying mechanisms. It also enables us to predict potential results under different policies, or in this case subsidising different products. This allows extrapolation beyond the specific RCT, examining for what types of health products and contexts would they expect the results to hold. The authors discuss other commonly subsidised health products: water purification products, cook stoves and deworming medicines might respond. By looking at the characteristics of these products and their demand, the authors are able to make predictions about how these products and the associated demand would be affected by subsidises. There can be differences in the level of accuracy in the quality of signals and in the disease reductions provided - some products might have negative short run side effects or the benefits may not be clear (i.e. deworming medicines have short-run negative side effects). Products differ also in durability (cooking stoves are much more durable than deworming medicine) and these

- The number of social contacts is set at c =0.4.

[9] See footnote 5 for full list of parameters that were set to specific values.

characteristics may limit the scope for longer-term learning. There may also be different health externalities for each type of product. It is extremely useful to be able to examine the implications as predicted by a validated model on other products. There is no need to run additional RCTs for each specific product in order to generate appropriate and well thought out predictions about the likely effects of different policies.

For instance, the authors conjecture that one-time subsidies for cook stoves and water filters have the potential to boost adoption via learning effects. The network health benefits are quite low for these products; thus the private returns remain large even if the neighbourhood uptake is high. In contrast, a one-time subsidy for water disinfectant is unlikely to have a meaningful impact on subsequent adoption, having the potential for negative learning effects due to its short run negative side effects and the lack of durability of the product.

The discussion of this case study shows how a simple model, can increase our understanding about the mechanisms of change, and allow the decomposition of the treatment effect into its various components. The RCT is complementary to the structural model, providing the exogeneity needed to identify effects and produce some reduced form results that will need to be explained. The experimental design was vital for the model, and the model helped to inform the design of the experiment in a way that would capture the differing channels. The complementarity of the two approaches are hopefully clear to the reader from our discussion of this example.

4.2 Case Study 2 : Incentive Work : Getting Teachers to Come to School (Duflo, Hanna & Ryan, 2012)

This study used a randomised experiment augmented with a structural model to examine whether monitoring and financial incentives could reduce teacher absence and subsequently increase learning in India. It is a good example for showing how structural models can test out a number of different policies to find the optimal one, without needing to run many different treatments. In India teacher absenteeism is a big problem, occurring about 25%-35% of the time, with low levels of academic competence in schools, over 65% of children enrolled in grades 2-5 are unable to read a simple paragraph. One hypothesis is that to increase performance, we need to also increase teacher presence. Financial incentives are one mechanism commonly used to improve presence, the added monetary benefit believed to increase labour supply. However, there might also be a negative effect if money crowds out intrinsic motivation to teach. Monitoring is another proposed mechanism to induce behavioural change.

In this RCT, teachers in 57 randomly selected Non Formal Education (NFE) schools were given a camera along with instructions to have children take photos of them and the other students at the start and end of every day. The camera's date/time settings could not be changed in order to prevent cheating. Each teacher was then paid according to the number of valid school days – with valid being defined as having the photos taken at least 5 hours apart and with 8 children in the photo.

They received Rs 500 if they attended less than 10 days in a given month, with an extra Rs 50 for each additional day attended that month. In the control schools, teachers were paid a fixed wage of Rs 1000.

The teachers were tracked for 30 months. At the end of the program, treatment schools had an absenteeism rate of 21%, with higher rates of 42% in the control schools and overall combined treatment and control baseline rates of 44%. Absenteeism was monitored via the photos, in addition to normal unanticipated visits to check on the teachers that were already conducted in the school system.

These results do not, however, reveal the potential effect of any other payment structure, nor of the contribution of daily monitoring data compared to the financial incentives effect.

In order to learn more about how this combination impacted absenteeism, the authors built a structural model of behaviour. They developed two different specifications of teacher labour supply – both model the dynamic decision process the teachers face as they accumulate days worked towards the bonus at the end of the month, both allow for unobserved heterogeneity at the teacher level. They differ in their treatment of the serial dependence in the opportunity cost of working. In the first model, the opportunity cost depends on whether the teacher attended work on the previous day. The second model posits that the opportunity cost is subject to an auto-correlated AR(1) shock. As it turned out, both models have similar results suggesting that the functional form chosen did not drive the results.

The experiment generates natural variation that can be used to identify the parameters of interest. The financial incentive for a teacher to attend school on a given day changes with the number of days previously worked that month and the numbers of days left in that month. This is because teachers must attend at least 10 days to receive any payment, and beyond that will earn an additional Rs. 50 per day worked. This payment structure also introduces a discontinuity when the month changes; this window is the source of variation used to identify the responsiveness to the bonus.

The dynamic model captures teachers' responses to the varying incentives over the month. Let *m* signify the month and *t* the day within the month where $t = \{1, \dots, T_m\}$. The teachers' utility function depends on their consumption and their leisure:

$$U_m = U(C_{tm}, L_{tm}) = \beta C_{tm}(\pi_{tm}) + (\mu_{tm} - P)L_{tm}$$

where *P* is the non-pecuniary cost of missing work. They assume linearity and separability in preferences. Consumption is a function of earned income (π_m). There is no discounting within a month, utility is linear and all income is consumed at the last day of each month, when they get paid. β converts consumption into utility terms. Leisure is equal to 1 when the teacher does not attend work, 0 otherwise.

The coefficient on leisure has the following form:

$$\mu_{tm} = \mu + \epsilon_{tm}$$

The deterministic part is the difference between value of leisure and the intrinsic value of teaching, the more they value teaching, the smaller is μ. The second term is a randomly generated shock.

Teachers who don't go to work face two penalties. They pay a cost *P*, they will be rebuked for not being present. If they take more absent days, they become more likely to get caught out in either system and reprimanded. There is also a possibility of being fired for absenteeism. The probability of being fired in any given period is $p_m(t, d)$, which depends on the number of days previously worked, by time *t* in month *m*. If fired, they receive a one-time payment *F.*

With *d* being the numbers of days worked in a given month, a teacher's income earned in the last period is given by:

$$\pi_m = 500 + 50.\max\{0, d_{m-1} - 10\}$$

The control group receive Rs 1000, independent of *d*.

All teachers are assumed to maximise the present value of lifetime utility.

The control group teachers face a repeated binary choice problem. The Bellman equation[10] for them every day of the month, except the last is:

$$V_m(t, d;\ \epsilon_m) = p_m(t, d).F + \left(1 - p_m(t, d)\right).\max\{\mu - P + \epsilon_{mt} + EV_m(t+1, d\ ;\ \epsilon_{mt+1}), EV_m(t+1, d+1\ ;\ \epsilon_{mt+1})\}$$

The current period utility of attending work is 0, as they only get a flat wage at the end of the period. Agents weigh the marginal change in the possibility over being fired in future periods against the immediate benefit of skipping work. The expectation over future value functions is taken with respect to the distribution of next periods' shock.

The authors redefine the outside option for the control group as $\tilde{\mu} = \mu - P$. At the end of each day, for $t < T_m$; t increases by 1 and d increases by 1 if they worked that day. After time T_m, everything resets to zero. On the last day of each month, the value function is almost identical to above with β x 1000 added to the utility of not being fired.

Those in the treatment group have a very different incentive structure. There is an additional dynamic component, teachers trade off immediate gratification with the possibility of increased wages at end of the month. There is also better information available to the authorities, with the data from the camera monitoring, leading to changes in both *P*, and the probability of being fired. The effect on *P* is unclear – it might increase due to the perfect information available from monitoring. But on the other hand, the monitoring body might still put more weight on an absence they find during a routine visit, and thus the expected cost of a missed day may in fact, be similar in both treatment and control. Or if the monitoring authorities feel that individuals get

[10] This is a commonly used technique in dynamic choice problems.

reprimanded enough via the financial loss, then *P* may go down. To acknowledge the non-pecuniary cost could be different it is denoted $\bar{P}$ for the treatment group.

Given the above, the payment structure for $t < T_m$ is:

$$V_m(t, d;\ \epsilon_m) = p_m(t, d).F + \left(1 - p_m(t, d)\right).\max\{\mu - \bar{P} + \epsilon_{mt} + EV_m(t+1, d\ ;\ \epsilon_{mt+1}), EV_m(t+1, d+1\ ;\ \epsilon_{mt+1})\}$$

At time T_m:

$$V_m(T_m, d;\ \epsilon_m) = p_m(T_m, d).F + \left(1 - p_m(T_m, d)\right).\max\{\mu - \bar{P} + \beta\pi(d) + EV_m(1, 0\ ;\ \epsilon_{mt+1}), EV_m(1, 0\ ;\ \epsilon_{mt+1})$$

To solve these value functions, the authors ignore any consideration of the next month in decision-making. This breaks an infinite time problem into a series of finite time problems.

Some points to note on identification and solutions:

As briefly mentioned previously, they estimate the model with a number of different specifications. First, a model with i.i.d. errors and then others with different forms of serial correlation (observed lagged absence, AR(1) process. The models can be extended to include observed and unobserved heterogeneity in the types of teachers. They incorporate observed characteristics into μ, such as the teachers' ability scores. They also allow for the outside option to vary using either fixed or random effects models.

The main parameters are calculated under different specifications, models with different forms of serial correlation, and models with extensions for observed and unobserved heterogeneity of the teachers. They calculate the implied labour supply elasticity: the percentage increase in the average number of days worked caused by a 1% increase in value of bonus and the semi-elasticity with respect to the bonus cut off. Of interest also is the percentage increase in the average number of days worked in response to an increase of one day in the minimum number of days needed for a bonus.

The results of the structural model show that teachers are responsive to the financial incentives, with an elasticity of labour supply with respect to the level of the bonus of about 0.2 – 0.3. This estimate is similar when using all the data, the first five days or a three-day window to identify the model. This suggests that identification based on the shift in incentives at the end of the month drives the results in all models and is thus a valid identification approach. In all of models, the mean outside option is negative for the majority of agents. This suggests that when considering the non-pecuniary cost of absence (*P*) without any financial incentives, teachers are willing to work more than half the days. When the financial incentive is set to zero, the predicted rate of absence is equal to that found in the control groups in the RCT, just under 50%. This suggests that $\bar{P}$ is close to P and perhaps there are no direct impacts of daily monitoring via the cameras. Maybe the current system of monitoring is enough to deter absenteeism (or perhaps this system is also redundant and there is some intrinsic reason why teachers turn up about half the time that is invariant to changes in the monitoring system).

The estimation only uses data from the treatment group. This allows the model to be tested by how well it predicts outcomes in the control group. If a model is a good fit then it should fit the control equally well.[11] The process of estimating the model using one part of the sample, and using this to predict other part of the data (holdout sample) is commonplace in machine learning and statistics. It is one way to check the appropriateness of the models that are being used. Another way is to see how well they predict other aspects of the data, such as predicted number of days worked.[12] In the simulated moments' estimation, this parameter was not designated as one to be matched, and as such it is a valid check on the fit of the model. The models tend to do reasonably well at predicting patterns in the control sample.

The model was then used to evaluate other possible policies that were not in the initial experimental design. This is one of the key benefits of any structural model and is especially clear in this case study. Here, the parameters are used to predict the optimal incentive scheme for any given number of days worked on average in a month. For example, if we want to incentivise the same number of days worked as the study above suggests (i.e. for an average of 17 days) the most effective way of doing so it to increase the bonus cut off to 21 days and the bonus to Rs 75 per day. This incentive structure provides savings of Rs 193 per teacher per month, a total of 22% of the total cost of the system tested. This is extremely useful for efficient policymakers – and one that is only possible through the use of a structural model.

There were no effects of the treatment on multi-tasking (a measure designed to see if teachers would work less when faced with financial incentives to turn up) which suggests that the marginal costs of teaching conditional on attendance was low. Student attendance was similar in both groups, so those with teachers receiving incentives got more instruction days. There were benefits on test scores, with a test score differential of 0.17SD in the treatment schools a year later; two and half years later there was a ten percentage point difference in likelihood of transferring to a formal primary school (which requires taking a test), between treatment and control groups.

This study shows that while building a structural model requires a lot of assumptions, there are ways to test the sensitivity of the results to its assumptions. This is good practise in econometrics: a good model should be robust to changing assumptions. It also shows various validation methods that can be used to test goodness of fit. A structural model has the flexibility of making predictions of optimal policies that are not part of the original RCT. The finding that the same level of reduced absenteeism could be obtained at lower cost is important and potentially useful – especially as it can be done without trialling each incentive scheme. This case study clearly shows the benefit of a structural model to test other policy options; whereas it would be near impossible to test more than a handful of options in a RCT setting.

---

[11] If the control group data had been used to estimate the model, then we cannot measure model fit in this way.

[12] Again, this can only be used in models that do not use the full dataset in estimation.

4.3 Case Study 3: Education Choices in Mexico: Using a Structural Model and a Randomised Experiment to Evaluate PROGRESA (Attanasio, Meghir & Santiago, 2011).

The final case study involves a more complex model than the previous two examples, but one that we believe is exemplary in showing the richness of the findings that is only possible with a structural model. It will also highlight that smart experimental design is vital to enable the identification of richer models.

Attanasio, Meghir & Santiago (2011) analysed the impact of monetary incentives on education choices in rural Mexico. This intervention was part of the PROGRESA program, a multi-faceted intervention started in 1998 by the Mexican government. Its main aim was to improve human capital accumulation in the poorest communities. The program provided cash transfers conditional on specific types of behaviour in three key areas: nutrition, health and education.

In the intervention analysed in this project, mothers in the poorest households in a set of targeted villages were given grants to keep their children in school in the final three years of primary school and for the first three years of secondary.

The authors estimated a structural model of education choices using the data from male children. They were able to incorporate general equilibrium and anticipation effects, to test for separability of preferences as well as for examining the impact of this particular intervention on school attendance. The authors were also able to suggest how to improve the programme based on simulations using different versions of the grant. PROGRESA changed the relative price of education and child labour in a controlled and exogenous fashion that enabled to identify the structural relationships within the proposed model. The discussion that follows explains that designing and/or using existing experiments to validate structural models is likely to lead to important advances in the understanding of this and other complex processes.

Using their model they found a number of key findings from the actual program.

1. The program has a positive reduced form effect on children's school enrolment, especially after primary school. This result is replicated in the structural model.
2. There were minimal anticipation effects of the grant. This important as it suggests that the difference between treatment and control is a valid comparison. If there had been anticipation effects then the control group might also have altered their behaviour, which would have resulted in an underestimation of the overall effect.
3. There was a general equilibrium effect on wages. The grant induced more children to go to school instead of working. This in turn raised the external wage, raising the opportunity cost of attending school, which attenuated some of the effects of the grant. However, the magnitude of the effect of the grant overwhelmed the effect on wages, which led to an overall positive effect.

One key advantage of structural model is the ability to evaluate programs that were not actually implemented and make useful policy recommendations based on these counterfactual exercises. The authors show that a revenue neutral change in the program that increases the grant for secondary school children while eliminating the grant for the primary school children would have a substantially larger effect on school enrolment rates of primary school children, and minor effects on the secondary school children, thereby increasing overall enrolment for the same cost.

Here we provide a brief outline of the components of the dynamic school participation model. This model is more complex than that of the previous two case studies, and we don't go into the full details of every step. Instead we will use it as an exposition of the richness of a full model of behaviour and of the many aspects of a model that potentially need to be considered. However, the many findings that are possible due to its richness, justify its complexity.

Each agent decides whether to attend school or to work, taking into account the economic incentives of each option. Education is assumed to end at age 18, and agents receive the reward of schooling (its' terminal value), a function of the highest grade passed. When the individual is in school they incur a cost, if they work they earn a wage. The grant is a monetary reward given to the individual for staying in school. The model incorporates uncertainty about the cost of schooling, grade progression, the wage process, and general equilibrium responses in the labour market. The model is dynamic; the child must continue in school today in order to complete a later grade. Second, the authors allow for state dependence; the number of years of schooling completed until today affects the utility of attending in this period.

There are costs and benefits of attending school. The utility function of a given agent in school is given by:

$$U_{it}^{s} = Y_{it}^{s} + \alpha g_{it}$$

$$Y_{it}^{s} = \mu_{i}^{s} + a^{s\prime} z_{it} + b^{s} ed_{it} + 1(p_{it} = 1)\beta^{p} x_{it}^{p} + 1(s_{it} = 1)\beta^{s} x_{it}^{s} + \varepsilon_{it}^{s}$$

where g_{it} is the amount of the grant an individual receives; which is zero for non eligible individuals and control villages. Y_{it}^{s} represents the costs/gains to attending school. The μ_{i}^{s} contains unobservables time constant features, the ed_{it} terms allows for the fact that attending school now has implications for future grades. z_{it} is a vector of taste shifts variables, including the normal parental background attributes and age and time dummies.

The first indicator $1(p_{it} = 1)$ denotes attendance in primary school, with $1(s_{it} = 1)$ denoting secondary school attendance. x_{it}^{j} represents the factors affecting the costs attending primary (j=p) and secondary (j=s) school. The error term is assumed to be i.i.d over time.

In comparison the utility of not attending school is:

$$U_{it}^{w} = Y_{it}^{w} + \delta w_{it}$$

$$Y_{it}^{w} = \mu_{i}^{w} + a^{w\prime} z_{it} + b^{w} ed_{it} + \varepsilon_{it}^{w}$$

where w_{it} are the potential earnings outside of school. The wage is a function of age and education, as well as location.

The model can only identify the differences between parameters that enter both equations. They can rewrite the preference equations as:

$$u_{it}^{s} = \gamma\delta g_{it} + \mu_i + a'z_{it} + b\, ed_{it} + 1(p_{it} = 1)\beta^{p}x_{it}^{p} + 1(s_{it} = 1)\beta^{s}x_{it}^{s} + \varepsilon_{it}$$

$$u_{it}^{w} = \delta w_{it}$$

where $a = a^{s} - a^{w}$, $b = b^{s} - b^{w}$, $\gamma = \alpha/\delta$, $\mu_i = \mu_i^{s} - \mu_i^{w}$ and $\varepsilon_{it} = \varepsilon_{it}^{s} - \varepsilon_{it}^{w}$.

The main coefficient of interest is γ; measuring the impact of the grant as a proportion of the impact of the wage on the investment in education decision. The grant is scaled so as to be comparable to the wage. If $\gamma = 1$, the wage and the grant have the same effect on education choices, which given the data is a testable implication.

The return to education is represented by the terminal value at age 18.

$$V(ed_{i,18}) = \frac{\alpha_1}{1 + \exp(-\alpha_2 * ed_{i,18})}$$

where $ed_{i,18}$is the education achieved at age 18. Schooling choices take into consideration both the current and future benefits of staying in school. Let $I \in [0,1]$ be the random increment to the grade that results from attending school at present. If successfully completed the grade $I = 1$, otherwise $I = 0$. There is a probability of success of $p_t^{s}(ed_{it})$. The value of attending school of someone who has completed ed_i years in school, of age t with characteristics z_{it} is:

$$V_{it}^{s}(ed_{it}|\Upsilon_{it})$$
$$= u_{it}^{s} + \beta\{[p_t^{s}(ed_{it} + 1)Emax[V_{it+1}^{s}(ed_{it} + 1), V_{it+1}^{w}(ed_{it} + 1)]$$
$$+ \big(1 - p_t^{s}(ed_{it} + 1)\big)Emax[V_{it+1}^{s}(ed_{it}), V_{it+1}^{w}(\, ed_{it})]\}$$

where the expectation is taken over the possible outcomes of the random shock, ε_{it}, and Υ_{it} is the entire set of variables known to the individual at period t , affecting preferences and expectations of costs and benefits of education and labour market opportunities. The value of working is:

$$V_{it}^{w}(ed_{it}|\Upsilon_{it}) = u_{it}^{w} + \beta\{Emax[V_{it+1}^{s}(ed_{it}), V_{it+1}^{w}(ed_{it})]\}$$

The difference between the first and second two terms of each equation reflects current costs of attending, whereas the difference between the second two terms reflects future benefits and costs. β is the discount factor, which incorporates liquidity constraints and other factors that lead households to disregard the future by differing degrees.

Wages are the opportunity cost of education. In the model an increase in wages will reduce school participation. Yet as wages are determined locally, the program may affect them because it may reduce labour supply of children. These general equilibrium effects will be larger if child labour is not substitutable with other forms of labour. To have a fully specified model the authors need to estimate a wage equation. Firstly, they

do not observe wages for those not working. Second, for the dynamic programming model to work, the individual must use a wage equation to predict their future wages. Finally, it is necessary to evaluate the effects of the program on wages). They specify a standard Mincer-type wage equation:

$$lnw_{ij} = q_j + a_1 age_i + a_2 educ_i + \varphi_{ij}$$

They assume education is exogenous for wages. In this scenario this appears to be a reasonable assumption, but one that is unlikely to hold in many other settings. This relationship between education and wages is extremely flat in the village, due to the limited opportunities for work in rural communities[13]. There are also no selection effects on wages based on participation. They estimate the wage equation using OLS and then use the predictions in the model.

The final aspect of the model that needs to be highlighted is how the authors address habits and initial conditions. The presence of ed_{it} creates a problem as the researchers do not observe the history of schooling for the children in the cross section. They cannot assume that μ_i is independent of past schooling decisions, decisions that are reflected in the current level of schooling, ed_{it}.

To circumvent this issue they estimate a reduced form for educational attainment up to the current date. They model the level of schooling as an ordered probit model with index function $h_i'\zeta + \xi\mu_i$.The vector h_i includes variables reflecting past schooling choices such as distance to secondary school in pre experimental years. Since school availability (measured for example with variables such as distance) changes over time, it can be used as an instrument in initial condition model that is excluded on subsequent school attendance choice, which depends on school availability only during the experiment. The probability of $ed_{it} = e$ and of child *i* attending school is:

$$P\left(ed_{it} = e, Attend_{it} = 1 \mid z_t, x_{it}^s, x_{it}^p, h_i, wage_{it}, \mu_i\right)$$
$$= P\left(ed_{it} = e \mid z_{it}, x_{it}^s, x_{it}^p, h_i, wage_{it}, \mu_i\right) * P\left(Attend_{it} = 1 \mid z_{it}, x_{it}^s, x_{it}^p, h_i, wage_{it}, \mu_i\right)$$

Variables that affect the lagged variability of schooling but not the current availability of schooling play a key role in identification of the model.

The exogenous variation in treatment and control villages, between eligible and non-eligible people, identifies the effect of the grant. These differences alone do not identify different amounts of the grant. However, the grant level differed by grade, and within each grade there were children of different ages. This provides an extra source of exogenous variation that can be used to identify the effect of the size of the grant.

The authors exploit the variation in the grant induced by the randomized experiment to estimate a more flexible specification than could be estimated without the program. The effect of the grant was allowed to be different to that of wages - even though variation in the conditional grant could have had the same effect as changes in child

[13] This means that wages do not respond to changes in education levels which results in a horizontal wage-labour supply equation.

wages, as they both modify the relative price of school versus work. It is equally plausible that the marginal utility of income differs depending on whether the child attends school or not. This is known as non-separability of preferences, here between consumption and schooling. If there are non-separable preferences then changes in the grant and in wages can have different effects on school enrolment. The experiment allows the authors to test the functional form of the preferences and, if necessary, to relax this separability restriction. This adds a new dimension to the discussion of the role of randomised experiments in designing policy. If separability is rejected, experimental variation is useful not only for the exogenous variation it induces but also because it helps identify economic effects that may otherwise be unidentifiable.

The structural model also allowed the identification of anticipation effects. PROGRESA created a control group by delaying the implementation of the program in some areas, rather than excluding them completely. It is therefore possible that the control villages react to the program prior to its implementation, depending on the degree to which they believe they will eventually receive it. A straight comparison between treatment and control areas may then underestimate the impact of the program, as the control group may have changed their behaviour in response to their anticipation of receiving the grant. A structural model that exploits other sources of variation, such as the variation of the grant with age, is then able to estimate the extent of anticipation effects.

The authors estimated the basic model using maximum likelihood estimation techniques, and data from the PROGRESA program, with a number of different specifications to test the robustness of their results to the underlying assumptions. Having estimated the effects, the authors ran simulations to see how well their models predict the impact of the grant and to quantify the effect of the main variables of interest. The model predicts an average effect of 0.047 on enrolment, similar to the reduced form finding of 0.05. It replicated the inverted U shape, in that the effects were largest for the children aged between 12-15 years old. The general equilibrium adjustment caused a decrease in labour supply of children and a subsequent increase in wages. The findings suggest that on average, reducing the wage by 44% (which is about the same size as the average grant received) increases the probability of attending school by 2.1%.

The structural model allows the authors to perform policy experiments by comparing the effect of the current program to one that differs in the way the grant varied by grade. This change was done to estimate the impact of a program that is targeted to those at the most responsive ages. They do this in a balanced budget approach so that the grant is increased for those above grade 6 and reduced for those below. The findings show that this change doubles the impact of the program. While this finding is unsurprising, given that most children in Mexico below grade 6 go to school anyway, it goes contrary to a lot of current government policies that tend to put their resources into younger years by default. The authors suggest that future policies to increase enrolment rates would benefit from modifying the age group targeted by the. From the households' perspective they receive the same amount of resources over time, but the timing of when they receive these now vary. If households can borrow against the future grant, then the effect of this reform would be to improve incentives for school participation at later ages. If however, families are liquidity constrained then this trade-

off may be more serious, particularly if the grant at a younger age improves nutrition or other child inputs. Clearly, there are a lot of moving parts that need to be considered when rolling out any policy.

The paper then considers a number of other policies such as decreasing the wage by an amount equivalent to the grant, changing the distance to school, and the impact of the grant on those with low cost of education. This shows the utility of a structural model verified using data from an experiment; again, we can go much further than the typical treatment - control difference. It helps to understand the mechanisms, and to answer more questions. It can be complicated and does require lots of moving parts but we believe that the richness of results justifies the costs involved in building a valid structural model of behaviour.

Section 5

Challenges in the Use of Structural Models

While the previous section presented case studies that show the benefit of using structural models with RCTs, we now discuss the challenges involved in fully implementing structural models into policy analysis. We believe the major drawbacks stem from the following issues: (1) the difficulty of tailoring structural parameters to relevant policy questions, (2) the lack of credible identification and (3) the use of untestable assumptions about behaviour.

First, it can be difficult to tailor the specific structural parameters being estimated to the policy questions of interest. Structural models need to be carefully specified, and certain behaviours - especially those that involve complex dynamics - can be difficult to capture accurately over time. The number of decisions and parameters can blow up when we turn our attention to life cycle models- and yet these dynamic models help to answer some of the most relevant policy questions in relation to skill accumulation, earnings and labour decisions, retirement and savings. Certainly, the increase in computing power over the past decades has helped the structural approach but it is still a difficult and time-consuming endeavour to undertake.

Second, even though randomised trials can help to identify the parameters, there will be cases when credible identification is not possible. There may be reasons why we cannot run a trial at all or in a way that supplies the necessary exogenous variation. A good example, for instance, is that of the use of mediation analysis, where a trial can be run to identify the impact of an intervention on a specific outcome but not necessarily to identify the mechanisms that lead to an observed outcome. In these cases, we would then normally rely on existing or observational datasets, of which we can try to find other sources of plausible exogenous variation to identify the model. However, even then the information may not be rich enough to be of used. The data may not be able to eliminate competing causal explanations of the same phenomena. The beliefs, attitudes and information that determine individual choices are often simply not observable in socio-economic surveys. This is not to say we can't use observational studies, we just have to think carefully about the source of randomness that we rely on for identification. And, in the design of surveys, we need to think about the construction of measurement tools that can help in solving identification problems. Situations may arise in which there is no data source that allows us to estimate the model and in this case it is futile to attempt to use this approach.

Third, while we have shown the benefit of the framework of structural models, the imposition of more structure on individual agents' decision problems is not always a good thing. Certainly, the assumptions made to identify and to estimate a given model need to be tested for validity, assessed against reality and relaxed where possible. The ideal situation is when a model is flexible enough to explain behaviour, and its predictions do not rely on any rigid and fundamental assumptions. It can sometimes be difficult to come up with genuine tests of the assumptions. Without such tests, however,

a structural model may fail the ultimate goal of any good theory: to be testable and to generate falsifiable predictions.

In a related vein, while some structural models are off-putting due to their complexity, there is equally a risk of building a model that is too simplistic. If one over simplifies the behaviour and economic environment, then using such a model to generate useful results and predictions will be a stretch. Striking the ground between overly simplistic and overly complex models can be tricky and is one of the reasons why building a good structural model can take a considerable time investment.

From a practical perspective, we recognise that structural modelling may be intimidating for researchers and practitioners alike. Researchers may be faced with a lack of resources, time and familiarity with the methods used. For structural modelling to become a common component in the apparatus of evaluation there will need to be a shift, not only in people's desire to use more rigorous modelling, but in how we think about evaluation in general. Currently, people tend to think within their own fields, their own expertise and look at what is normally done in their field. How do we go about changing these norms so that structural modelling becomes the norm? We believe this requires top down and bottom up change. CEDIL and its work with DFID provide an opportunity to demand more rigorous evaluation methods from contractors, but we recognise that many researchers will need support in learning these methods. This could involve specific training and guidance provided by economists, or to enlist the help of someone who is trained in these techniques. People will need to be convinced of the utility of these methods, but also of the credibility of the findings and their applicability to evaluation questions.

Section 6

Conclusion

This paper shows the utility and richness of results that can be uncovered when using structural models. The authors believe that structural modelling, when used in conjunction with other research techniques that policymakers are already familiar with (for examples randomised control trials) can be a valuable methodology to further our understanding of many problems. Behavioural models need not be theoretically or mathematically complex, and researchers should not be put off by a stereotypical and oftentimes unfair view of structural models as imposing untestable assumptions and restrictive structure. Of course, implementing a structural model requires a clear understanding of the decision problem, and can require a significant time and cognitive investment on the part of the researcher. However, all quality research needs to be done with careful consideration of the specific policy, the mechanisms through which it could be working, and the cases in which the results will or will not hold. Structural modelling forces you to think about these elements and the payoff (in terms of richer results) is, in our opinion, worth it. We hope that structural models will become a standard technique in policy evaluation, being used to help answer a wide range of policy questions.

Implications for CEDIL and DFID

The Centre for Excellence in Development Impact and Learning was set up to inform DFIDs' strategic plan over the coming decade with the aim of the successful attainment of the sustainable development goals. It hopes to effect a step change in the rigour and usefulness of evaluation in international development.

We believe that structural modelling and its use together with other approaches can be a key mechanism through which this step change can be achieved. Structural modelling is a rigorous approach to evaluation, generating useful and widely applicable findings that can be used in policy design. It also offers a way to widen the scope of questions that can be answered and policies that can be trialled.

CEDIL identified four major gaps that it wanted to address. These were: (1) thematic and geographical gaps, (2) method gaps for evaluating primary studies, (3) synthesis gaps and (4) evidence translation gaps. Structural modelling can help in each of these.

For thematic gaps, structural modelling can address questions that are not answerable with normal research RCT methods for several reasons. While most of the examples in this paper have come from education and health, structural modelling has been used to evaluate large scale infrastructure projects – such as the benefits of railroad system in India on economic growth by Donaldson (2010). It has also been used in resource and environmental economics – an interesting review paper by Timmins and Schlenker (2009) discusses how reduced form and structural models have been used. Structural models can be applied in many realms that are typically 'difficult' to evaluate.

In regard to synthesis gaps, structural models offer a method in which the results of many different RCTs or question can be assessed and weaved in a coherent narrative. The behavioural models used should be such that they extends to different policies and intervention that try to influence similar behaviours. A coherent structural model allows an organising framework through which a question can be examined. For evidence translation gaps, we believe that structural modelling, while complex, offers a useful framework to understand behaviour. By being explicit about the channels of change, policy makers will hopefully gain a deeper understanding of how and why the policy is working, when it won't work or what might impede its success. It is believed that it can, if presented in an accessible manner, actually help the consumer of policy research understand what is going on and thus help inform future policy uptake.

It is perhaps the second gap, of evaluation of primary studies that will gain the most from the use of structural modelling. Structural modelling is not a new research tool; it is used and has been used by many researchers in Economics, and perhaps under different names in other fields. We hope to persuade people who have not heard of it, or remain sceptical of it, that it should become a normal evaluation technique, used in conjunction with existing methods (most obviously with RCTs). In an ideal world the process of thinking through a structural model of behaviour would help to inform the design of the randomised control trial and what data is collected. However, as an immediate step we could identify some high quality RCTs that may be amenable to

identify a structural model. This could allow researchers to start work immediately and to increase the evidence around the gains in utility from combining structure to RCTs.

Specifically, for the programmes of work we would recommend that CEDIL should promote an integrated structural model with RCT approach. The tender would ideally involve a large-scale flagship programme, like PROGRESA, as the combined approach can be quite resource intensive. It could also be used to promote a project that could build a structural model to improve our understanding of certain problems and make future predictions of the impact of interventions, using existing data, either from an RCT or via observational datasets with a clear identification strategy. This would allow immediate use of the methods, as opposed to waiting for an RCT to be conducted. Of course, this would require the correct identification of a suitable project. Future programmes of work make it mandatory to include a structural model, to estimate the effects and to look at mechanisms.

CEDIL could also offer training workshops as part of their goal to improve knowledge sharing between disciplines. A basic crash course in understanding, building and estimating structural models and how these can be applied and used in impact studies could be offered to participants. There may be merit in potentially offering two different versions of a similar course, depending on the strength of quantitative skills the researcher/policy maker has.

For CEDIL to fulfil its mission of becoming an *"internationally recognised leader in methods developed for production and synthesis of rigorous evidence of what works, why, where, for whom and at what cost"*, we believe that the Centre should require researchers to go beyond simple treatment-control evaluations and to think more deeply about how best to address the problem under examination. Structural modelling is one method through which more rigorous evaluation can be achieved. DFID and CEDIL can benefit from incorporating structural modelling, or at least requiring individual evaluators to think more deeply about the mechanisms of change, the scalability and generalizability, into its requirements of any future evaluations.

References

Athey S and Imbens GW (2017) The state of applied econometrics: causality and policyevaluation. *The Journal of Economic Perspectives* 31(2): 3-32.

Attanasio O., Cattan S., Fitzsimons E., Meghir C. and M. Rubio-Codina (2015). 'Estimating the production function for human capital: Results from a randomized control trial in Colombia. *National Bureau of Economic Research*, Working Paper No. 20965.

Attanasio, O, Levell, P, Low, H and Sanchez-Marcos V (2018): Aggregating Elasticities: Intensive and Extensive Margins of Female Labour Supply *National Bureau of Economic Research*, Working Paper No. 21315, forthcoming in *Econometrica.*

Attanasio O, Meghir C and Santiago A (2011) 'Education choices in Mexico: using a structural model and a randomized experiment to evaluate Progresa.' *The Review of Economic Studies* 79(1): 37-66.

Attanasio O. and E. Pastorino (2018): 'Non-linear prices in village economies', NBER WP 21718.

Bollen KA and Pearl J (2013) Eight myths about causality and structural equation models. In *Handbook of causal analysis for social research*. Springer, Dordrecht, pp.301-328.

Browning M (2012) Two examples of structural modelling. Notes for "Structural Modelling'. Department of Economics, University of Oxford. Available from: https://pdfs.semanticscholar.org/651d/bff4f4895001b39356615c1481de8321caa7.pdf

Carneiro P, Heckman, JJ and Vytlacil E (2010) Evaluating marginal policy changes and the average effect of treatment for individuals at the margin. *Econometrica* 78(1): 377-394.

Chetty R (2009) Sufficient statistics for welfare analysis: A bridge between structural and reduced-form methods. *Annual Review of Economics* 1(1): 451-488.

Deaton A and Cartwright N (2016) Understanding and misunderstanding randomized controlled trials. *National Bureau of Economic Research*. Working Paper No. 22595.

Donaldson, D. (2018) Estimating the Impact of Transportation Infrastructure. American Economic Review, Forthcoming.

Duflo E, Hanna R and Ryan SP (2012) Incentives work: getting teachers to come to school. *American Economic Review* 102(4): 1241-1278.

Dupas P (2014) Short-run subsidies and long-run adoption of new health products: evidence from a field experiment. *Econometrica* 82(1): 197-228.

Felstein MS (1999) Tax avoidance and the deadweight loss of the income tax. *Review of Economics and Statistics* 81(4): 673-680.

Foster EM (2010) Causal Inference and Developmental Psychology. *Developmental Psychology* 46(6): 1454–1480.

Goldberger AS (1972) Structural Equation Methods in the Social Sciences. *Econometrica* 40(6): 979-1001.

Hawley, W. A., P. A. Phillips-Howard, F. O. ter Kuile, et al. (2003). Community-wide effects of permethrin-treated bed nets on child mortality and malaria morbidity in western Kenya. *American Journal of Tropical Medicine and Hygiene* 68 (4 Suppl), 121–127.

Heckman JJ, Pinto, R and Savelyev, P (2013) Understanding the Mechanisms Through Which an Influential Early Childhood Program Boosted Adult Outcomes. American Economic Review 103(6): 2052-2086.

Heckman JJ and Urzua S (2010) Comparing IV with structural models: what simple IV can and cannot identify. *Journal of Econometrics* 156(1): 27-37.

Heckman JJ and Vytlacil EJ (2007) Econometric evaluation of social programs, part I: causal models, structural models and econometric policy evaluation. *Handbook of Econometrics* 6:4779-4874.

Heckman JJ and Vytlacil E (2005) Structural equations, treatment effects, and econometric policy evaluation. *Econometrica* 73(3): 669-738.

Heckman JJ (2000) Causal parameters and policy analysis in economics: a twentieth century retrospective. *The Quarterly Journal of Economics* 115(1): 45-97.

Hoddinott, J., E. Skoufias, and R. Washburn (2000): "The Impact of Progresa on Consumption: A Final Report," International Food Policy Research Institute, available at https://www.ifpri.org/publication/- impact-progresa-consumption.

Imai K, Tingley D and Keele O (2010) A general approach to casual mediation analysis. *Psychological Methods* 15(4): 309-334.

Killeen, G. F., T. A. Smith, H. M. Ferguson, et al. (2007). Preventing childhood malaria in Africa by protecting adults from mosquitoes with insecticide-treated nets. *PLoS Medicine* 4 (7).

Keane M (2010) Structural vs. atheoretic approaches to econometrics. *Journal of Econometrics* 156(1): 3-20.

Keele L, Tingley D and Yamamoto T (2015) Identifying mechanisms behind policy interventions via causal mediation analysis. *Journal of Policy Analysis and Management* 34(4): 937-963.

Kline P and Walters CR (2016) Evaluating public programs with close substitutes : the case of head start. *The Quarterly Journal of Economics* 131(4): 1795-1848.

Pearl J (1995) Causal diagrams for empirical research. *Biometrika* 82(4): 669-710.

Pearl J (2014) Interpretation and identification of causal mediation. *Psychological Methods* 19(4): 459-481.

Wright S (1921) Correlation and causation. *Journal of agricultural research 20*(7): 557-585.

Timmins C and Schlenker W (2009) Reduced-form versus structural modelling in environmental and resource economics. *Annual Review of Resource Economics* 1(1): 351-380.

Wooldridge J (2010) *Introductory Econometrics: A Modern Approach.* South-Western College Publishing.

Contact Us

Centre of Excellence for Development Impact and Learning (CEDIL)
London International Development Centre
36 Gordon Square
London WC1H 0PD
United Kingdom
www.cedilprogramme.org
@CEDILProgramme