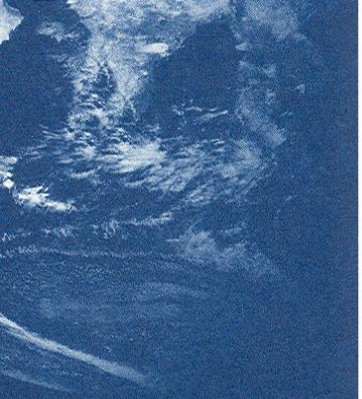

MAASTRICHT GRADUATE SCHOOL of GOVERNANCE

Evaluating causal effects of Early Childhood Care and Education Investments: A discussion of the researcher's toolkit

Robert Bauchmüller

Working paper
MGSoG/2008/WP006
October 2008

www.governance.unimaas.nl

Maastricht Graduate School of Governance

The 'watch dog' role of the media, the impact of migration processes, health care access for children in developing countries, mitigation of the effects of Global Warming are typical examples of governance issues – issues to be tackled at the base; issues to be solved by creating and implementing effective policy.

The Maastricht Graduate School of Governance, Maastricht University, prepares students to pave the road for innovative policy developments in Europe and the world today.

Our master's and PhD programmes train you in analysing, monitoring and evaluating public policy in order to strengthen democratic governance in domestic and international organisations. The School carefully crafts its training activities to give national and international organisations, scholars and professionals the tools needed to harness the strengths of changing organisations and solve today's challenges, and more importantly, the ones of tomorrow.

Author
Robert Bauchmüller, PhD fellow
Maastricht Graduate School of Governance
Maastricht University
Email: Robert.bauchmuller@governance.unimaas.nl

Mailing address
Universiteit Maastricht
Maastricht Graduate School of Governance
P.O. Box 616
6200 MD Maastricht
The Netherlands

Visiting address
Kapoenstraat 2, 6211 KW Maastricht
Phone: +31 43 3884650
Fax: +31 43 3884864
Email: info-gov@governance.unimaas.nl

- Abstract-

Evaluations of causal effects of early childhood care and education (*ECCE*) investments on later child outcomes usually encounter a number of identification problems. This is due to a particularly difficult quest to identify the counterfactual case. Solutions for this problem are provided by a number of identification strategies which are used in programme evaluation. In this paper, these identification strategies are explained and explicitly studied with respect to their applications in evaluations of *ECCE* investments. The discussion of this toolkit of strategies reveals that the availability of data determines which strategy should be preferred. The best strategy should build on theoretical foundations, mention the assumptions which are made when simplifying reality and should refer to the internal and external validity of inferences that are made. *ECCE* evaluations that are done in is this way, can better inform policymakers about what investments work best for young children.

Keywords: programme evaluation, causal inferences, early childhood care and education (*ECCE*), identification strategies

Acknowledgement: *I would like to thank the European Centre for Analysis in the Social Sciences which is funded by the European Community - Research Infrastructure Action under the FP6 "Structuring the European Research Area" Programme. The centre has supported my research visit to the Institute for Social & Economic Research at the University of Essex in the summer 2007. During this visit, could discuss with researchers at the institute the methods discussed in this paper and their applications to evaluations of early childhood care and education investments*

INTRODUCTION

Providing policymakers with evidence on the effectiveness of social programmes is an eventual goal of researchers in social sciences. The essential challenge of their research is to construct a research design in which true causal effects can be distinguished from mere predictive correlations. Among the disciplines of social science, economists have shown the strongest commitment to a distinction between predictive associations and causal effects given their principal interest in individuals' choices which are a major unobserved source of biased estimates. The concept of causal inferences goes back to the early work on experimentation, for example, of Fisher (1918) and Cochran (1965) and has later been developed into the theory of causal inferences[1] as espoused by Rubin (1974), Rubin and Rosenbaum (1983) and Holland (1986). A true causal effect is a measurable change in an outcome that results from exposing the unit of analysis to a treatment[2]. For example, participation in a social programme such as early childhood care and education can change the development of a child as measured by educational achievements. To observe true causality this rationality needs to be valid in the counterfactual case, too – if all other factors are held constant and the same unit was not treated, then there would not be that outcome[3]. Those who do receive the treatment are referred to as treatment group, those who do not receive the treatment are called control group. A well-formulated causal inference refers to a treatment which is potentially reversible. As argued by Imbens and Rubin, sex and race, for instance, are not easily reversible, thus statements like "the child had better language skills because it was female" would be an ill-formulated causal inference (Holland, 1986; Imbens & Rubin, 2008; chapter 1). This is a per unit definition of causal effects. In reality outcomes of only one state of treatment can be observed per unit; the counterfactual case cannot be observed (Holland, 1986) – the problem of causal inference is a problem of missing data. Therefore causal inferences rely on comparisons across distinct units, either of the same physical unit across different moments of time as,

[1] This theory is sometimes referred to as the Rubin-Rosenbaum-Holland (RRH) theory of causal inferences (see e.g. Shonkoff & Phillips, 2000; appendix B).

[2] The terms 'intervention', 'manipulation', 'programme participation' and 'action' are used interchangeably for the term 'treatment'.

[3] The reverse argument is a causal inference as well. If the treatment would be defined as the passive one – not participating in the social programme – and the outcome of this non-participation would be observed, then the counterfactual case would be the outcome if those non-participants would potentially have participated in the programme.

for instance, observed in longitudinal studies, or alternatively across different physical units at the same moment of time as, for instance, in cross-sectional studies.

The arena of early childhood care and education policies, including related family policies, has gained increasing importance due to changes in the economic, demographic and social composition of modern societies. More women stay in the education system, for longer. Female employment has risen substantially, demanding new modes of childcare. Family patterns are changing, leading to new arrangements of care for the children. Fertility has decreased in most developed countries. Early childhood is increasingly acknowledged as being a critical period for brain and cognitive development and a sensitive period for non-cognitive and language development, building the foundation for future skill development and achievements throughout the child's lifecycle[4]. Economists have argued that early investment in early childhood has two intrinsic advantages which make it the human capital investment with the highest private and social returns: 1) dynamic self-productivity implies that early developmental gains in domains such as motivation and curiosity enhance the interest in learning at later ages, and 2) multiplier effects imply that abilities which are learnt in this period enhance future skills accomplishments (Cunha, Heckman, Lochner, & Masterov, 2005; Knudsen, Heckman, Cameron, & Shonkoff, 2006; Shonkoff & Phillips, 2000).

The research agenda on outcomes of Early Childhood Care and Education (*ECCE* henceforth) investments applies different theories to make abstract representations of the empirical descriptions, for example, on segregation and socialisation processes (sociologists), on child development aspects such as learning (psychologists), on physical aspects such as brain maturation (neuroscientists) or human capital investments (economists). Economists are concerned with analysing the determinants of early human capital accumulation, which reflects future productivity of the person in the labour market (see, for instance, Becker, 1964) and ways to make a person more successful in

[4] Critical periods are those periods which are indispensable for developments of certain skills – a certain treatment or event must occur in this period so that development can proceed; sensitive periods are times in which skills can change significantly but are not crucial for later development – a certain treatment of event can cause an effect in this period.

the labour market (e.g.: Schultz, 1961). Economists also study the dynamics of human capital development throughout the lifetime and a wide range of factors that might determine the stock and the development of human capital. Examples of such factors are investments in out-of-family education, as for instance at day care centres and schools, learning experiences within the family as well as migration, health, job experience etc. Accordingly, *ECCE* investments are seen potential ways to improve the long-term labour market performance already at the stage of early childhood. While Becker and Tomes (1979) considered that such early investments might be substitutes for investments at later stages and looked at childhood as one static developmental stage, Cunha and Heckman stressed the dynamic character of investments at early stages. Early investments can enhance later human capital development both through being complementary for later investments and through causing self-productive spill-over effects (Cunha & Heckman, 2007a, 2007b). Heckman stresses the need to have profound theoretical hypotheses as starting point of an empirical analysis (Heckman & Vytlacil, 2007). One way to formulise theoretical hypothesis is the use of a mathematic model. Independent of what kind of model is going to be tested, researchers need to state clearly the abstraction of their analysis, for instance, by stating the assumptions made. Given such clarity statistical inferences can more easily be understood by other researchers.

The above mentioned trends have recently reactivated social policy debates on this topic, which lead to many new and reformed social programmes dealing with *ECCE*. At the same time, policymaking today is more prone to an evaluation culture that asks for hard evidence on outcomes of public investments ('hard' meaning quantitative evidence from statistical-empirical studies). Some qualitative studies provide valuable insights into the circumstances of *ECCE* investments and the wide variation of characteristics of individual children and their families. However, they are usually limited in terms of being valid in other (external) contexts. Policymakers acknowledge the fact that early childhood care and education policies are still ill informed in terms of hard evidence. More research on this topic is being promoted, for example, by providing more funds for large-scale cross-country comparisons, longitudinal studies and sophisticated empirical policy evaluations. Policymakers are interested in answers to policy questions such as: Which

early childhood care and education investments are most effective in promoting early childhood development? Whether and when can early maternal employment be detrimental for the development of the child? Which investments in non-parental care can promote child development? Which investments compensate less stimulating home environments? What impact do changes in family structures, parental labour supply and gender roles have on children's development? Are there any programmes that are more beneficial for some groups than for others? This is a selection of the predominant research questions. Theory would say that those *ECCE* investments which are particularly targeted on the individual, disadvantaged child, which are most intensive as well as consistent and which are continued through to later investments are most likely to change the child's early development and in turn bring about the greatest benefits for later outcomes. Despite empirical testing of those theoretical expectations 'on children who are most at risk', it is of interest to understand how the results generalise to children across the full range of individual risks. Various scientific disciplines have been studying such matters; this interdisciplinary cooperation has enriched our understanding of the impact of early investments in child development.

Usually *ECCE* programmes refer to the period starting at the end of parental leave and continuing until entering compulsory education at primary school age. Sometimes this conceptual framing is also extended to include the pre- and immediate post-natal period, for instance as well as additional care accompanying later compulsory schooling. Given the afore-mentioned policy questions, researchers typically look at the individual child as a unit of analysis. Aggregated units such as classes, schools or national systems may be used, for example, in comparative studies. The great variety of assessed *ECCE* treatments can be simplified into seven categories[5]: 1) pre and post-natal environmental, (health) care and nutrition inputs, 2) employment and leave patterns of both parents, and the related parental time investments in childcare, 3) attendance of diverse early non-parental childcare arrangements, 4) participation in early educational programmes such as preschool or kindergarten that are considered to actively promote various aspects of early

[5] The following studies provide a valuable overview of the main conclusions drawn from evaluations of ECCE investments: (Cunha, Heckman, Lochner, & Masterov, 2005; Heckman, 2008b; Karoly, Kilburn, & Cannon, 2005; Waldfogel, Han, & Brooks-Gunn, 2002).

child development and hence the child's school readiness, 5) neighbourhood and community activation programmes, 6) home visiting and parent counselling programmes, as well as 7) care and therapy following potentially traumatizing experiences during early childhood such as parental divorce, a death, accident or severe illness of the child or a close family member, or sudden impoverishment. Sometimes programmes incorporate two or more of these programme categories. Examples are programmes at which two-generation, parents and children, are targeted jointly. However, evaluation studies usually refer to single treatment effects in which only one of those investments is assessed. Nonetheless, the researcher's ultimate goal is to make more sophisticated claims (see e.g. Rutter, 2002): How do those treatment effects differ with respect to programme features? How do they differ with respect to their targeting and outreach? Which aspect of the multidimensional development domains is studied? What is the impact of environmental circumstances on the treatment effects? And, how do those aspects interact with each other?

Researchers want to open the black box of variations in the individual treatments as much as possible. The treatment, as proscribed by the programme's officially communicated rules, may not be the same as the one which an individual actually receives. Any indication of differentiation, for example, in categorical or ordinal sub-classifications improves the analysis. In this regard, efforts to differentiate between different treatments are, for instance, to distinguish different types of care, for example, centre and family based care; developmental and non-developmental time investments; nutritional, health or pedagogic care; or they might be characterized by more detailed quality characteristics such as child-staff ratios and training levels of staff. The treatments can also be distinguished in terms of doses of use, for example, in terms of hours per week and number of years attended. While more sophisticated data might enhance differentiation, such studies remain simplifications of reality as they still consider individual programmes or their aspects. When doing causal inferences for only a part of reality, two implicit simplifying assumptions are made: a) units are not interfering with each other and b)

treatments do not differ in terms of doses and qualities that all the treated receive[6]. The ultimate goal of researchers is to be able to do general analyses that acknowledge the various facets of such programmes while at the same time taking into account the various interactions between programme aspects and the characteristics of the participating children and their families. An example of interference is the limitation of parental time for different siblings. If the assumption of non-interference does not hold, additional information, for example, on parental time constraints, needs to be added to the estimation.

In studying the causal effects of *ECCE* investments, a wide variety of child outcomes, reflecting the stock of human capital, has been observed. A classic distinction looks at cognitive and non-cognitive abilities, while some studies also looking at health status. Cognitive abilities are reflected in skills that mirror knowledge and understanding, for instance, of how to communicate, write, read and do mathematical calculations. Non-cognitive abilities refer to the emotional and behavioural development which is reflected in measures of personality traits, such as motivation, self-control, perseverance, dependability, consistency, self esteem and optimism, but also in behavioural outcomes, such as delinquent and crime behaviour. Economic research has traditionally focused more on the analysis of cognitive skills. Recent debates acknowledge that the division of cognitive and non-cognitive skills is too artificial, that abilities are multi-dimensional in their character. However, the interaction between cognitive and non-cognitive aspects is not yet fully understood. What is known is that the malleability of cognitive abilities decreases with age, while non-cognitive skills stay more malleable over the life-cycle. Economic research has long been agnostic of the use of psychological test scores as economists had not fully understood the underlying traits, but also because they have been sceptical about the predictive power of such test scores on later outcome measures (see for example: Card & Krueger, 1996). In contrast, Herrnstein and Murray showed in their controversial book 'The Bell curve' (1994) that cognitive skills, measured as IQ in

[6] This assumption is referred to for instance by Imbens and Rubin (2008; chapter 1) as the Stable Unit Treatment Value Assumption (SUTVA). See (Heckman, Lochner, & Taber, 1998) for a discussion of this assumption in the case of large scale programmes and how potential interferences across units can be accounted for.

adolescence can predict later educational attainment and socio-economic status. Borghans, Duckworth, Heckman and ter Weel (2008) outlined that non-cognitive skills as, for example, documented in personality traits would have similar predictive powers on later outcomes as cognitive skills have.

Sometimes latent factors, such as test outcomes, cannot be directly observed; hence they are constructed from other observed variables[7] by means of clustering techniques, such as e.g. factor analysis, item response analysis or Rasch modelling. An issue of discussion on the use of a latent variable is whether the correct construct system can be identified. The construct system refers to the relationship between the variables and the latent variable. The question is then whether variable used as latent variable is actually related to the real latent factor that shall be measured – 'it is not an actual but a mental relation, revealing the structure of the theories rather than a structure in reality' (Borsboom, Mellenbergh, & Heerden, 2003). The creation of the best latent variables for personality traits is a good example on how much such construct systems can actually be subject for debate. Depending on the structure of the personality theory applied, many different latent variables systems could be constructed. While the majority of psychological studies tends to use the big five system of Openness, Consciousness, Extraversion, Agreeableness and Neuroticism (OCEAN), some argue in favour of using a more extensive set of latent variables to cover personality while others stress that two or three factors would already be sufficient. Borghans et al. (2008) discussed the economic interpretation and applicability of personality traits as measured in various tests by psychologists and, in reverse, the use of economic theory and methods to improve the use of tests scores by psychologists. They point out that the creation of indicators for personality traits needs to be guided by strong theoretical foundations. Reverse causalities between latent factors and the used child outcomes as well as environmental contexts and response incentives need to be carefully accounted for.

Another important differentiation in child outcomes is to distinguish short-term from long-term outcomes. Short-term outcomes are observed within or immediately after the

[7] The observed variables that are used to construct a latent variable are also called 'manifest' variables.

early childhood period and long-term outcomes are observed in later childhood and early adulthood, for example, after the age of 16. Short-term outcomes are usually seen in terms of achievements, whether educational attainments within the school career - reflecting the capacity to master a certain school curriculum within a distinctive class environment, or as developmental tests - reflecting innate abilities at certain ages. However, the research interest lies eventually with longer term outcomes, which are more predictive of lifetime opportunities and intergenerational mobility. These two aspects underlie policymakers' attempt to attain equal opportunity and full social mobility for all. Later educational attainments at the end of secondary schooling or at tertiary education and early labour market performance are good predictors for later outcomes; which can be measured, for example, in terms of income from labour or the propensity to depend on social transfers, to live healthier or to take part in criminal activities

Despite of wanting to know the, Policymakers are not only interested in the overall causal effect on child outcomes but also differences between social groups. Likewise, researchers are not only interested in the main effects on the child outcome levels per se, but also in the distribution of those effects across the whole population. This is particularly true for those outcomes which cannot be corrected for easily by redistributive social welfare transfers. Heckman pointed out that later remedial interventions, to correct for development deficiencies, are rather ineffective (Cunha & Heckman, 2007b; Heckman, Krueger, & Friedman, 2002). Human capital gains such as those from *ECCE* investments cannot be transferred between people. However, according to Heckman's synergy model, early gains in skill accumulation can be transferred dynamically to later stages in life; it is, therefore, of interest to understand how the distribution of skills can be influenced within *ECCE* investments at the first place. Accordingly, sociological research into intergenerational mobility tends to differentiate groups along socio-cultural, ethnic and occupational lines whereas economic research tends to differentiate across different incomes and performances in the labour market. Seminal works such as Coleman's report (1966), which turned attention from equality of inputs to equality of outcomes have turned the research focus (back) towards understanding the full range of environmental

and family background characteristics[8]. In this regard, research on the effects of *ECCE* investments is still incomplete in covering those characteristics. Heckman, Urzua and Vytlacil (2006), consequently, point out that evaluation strategies need to address such sources of heterogeneity in treatment effects carefully.

Choosing which child outcome(s) will be the focus of the research requires as much care as the specification of the treatment that is evaluated. Despite being explicit in inferring which treatment causes which outcome, the need to simplify the reality for the empirical analysis demands clarity about which aspect of child development is studied with the chosen child outcome variable. Being explicit about which treatment and which child outcomes are studied facilitates their comparison across studies and different settings. In making claims about the relationship between cause and effect, it is essential to assess whether those claims can be made in general for every situation - ceteribus paribus (all else being equal) - or whether those claims can only be made for the observed units that are evaluated. Many studies of *ECCE* effects refer to small samples which are not easily comparable to children and their *ECCE* and family settings in other countries. Even if such studies are internally valid in terms of the chosen identification strategy for drawing causal inferences on the studied sample, the validity of inferences in other contexts and historical periods might still be limited. This is particularly true if the sample size is too small or the studied environmental setting is too different from other settings as it is therefore representative. Often policymakers are satisfied if the external validity of causal inferences is guaranteed according to a certain local setting and hence their own sphere of political activity. However, for the scientific researcher it is of more importance to make inferences with stronger external validity. They strive to reach external validity in order to enhance the full understanding of the interaction of local contexts with studied treatments.

[8] Other influential works, acknowledging the importance of environmental and family background factors as compared to the influence of schooling and income factors have been Jencks' study on 'Who get ahead?' (1979), Mayer's study on 'What Money Can't Buy' (1997) as well as Bowles, Gintis and Osborne Groves' study on 'Unequal Chances' (2005).

There has been an increasing interest in this research area, which has led to an extensive data collection. Various cohort studies follow children, starting at birth or even during pregnancy, and include a wide variety of information about the early childhood period and different developmental outcomes. The Millennium Cohort Study (MCS), for example, started closely following a sample of British children born in the year 2000 and the Longitudinal Study of Australian Children (LSAC) has followed children born in the year 2000 and 2002. However, despite the improved availability of data, there will always be a limitation in terms of complete data on all aspects of reality. Theory and earlier findings can be a guide in terms of which information shall be collected. As long as limited data is used, statistical methods are needed to make causal inferences. In social sciences researchers tend not to be able to prove that a causal relationship actually exists; using statistical methods they can only exclude - with a certain probability - the option that a causal relationship is non-existent. Empirical studies on the effects of *ECCE* investments on child outcomes are often challenged by econometricians and economists as being problematic in terms of their 1) too loose interpretation of the statistical results of empirical estimations, their 2) limited theoretical foundations or their 3) imprecise problem identification. In other words, they are concerned that the available data is too limited to give extensive conclusions about causal relationships, in particular if several relationships are studied at the same time. They are also concerned about whether theory has guided the empirical testing of hypotheses or whether empirical tests have resulted in formulating hypotheses about relationships ex-post rather than ex-ante of the empirical investigation. Yet, a major concern is whether the researcher knows enough about the population under study, in particular about the underlying sampling processes, the treatment assignment and the relevance of unobserved characteristics, so as to be able to make causal inferences. Moreover, interpretations of statistical outcomes are often influenced by differing views on the population, which leads to normative statements (Manski, 1993). In identifying the research strategy for a policy evaluation it is essential to state the extent to which reality will be simplified. This implies explicitly stating the assumptions the researcher makes about reality and defending how plausible those assumptions actually are. Using this approach, social scientists from different disciplines

could debate the assumptions used for an investigation rather than the validity of conclusions of an entire discipline in general.

There are three major concerns regarding evaluations of investments of care and education for children in their early childhood: 1) endogeneity biases, 2) simultaneity biases, and 3) selection biases. Are the estimation results *endogenously biased* by unobserved heterogeneity in the children's endowments? Were for example the children who benefited from an *ECCE* programme initially different from the children who were not treated, for example, in terms of their level of development at the beginning of the programme? If such differences are not account for, positive outcome might be wrongly attributed, for instance, to an early education programme when actually the positive outcomes could be related to the participating children already being more advanced than the control group, before the programme starts. Accordingly, there might be a reason of concern if parents of those children who had early developmental drawbacks adapted their parenting style. While the child might participate in an *ECCE* investment, it is difficult to causally differentiate between the effect of the *ECCE* investment and the *simultaneous adaptive behaviour* of parents. Similarly, economists question whether estimation results correctly accounted for the choices that were made regarding participation in the programme, for instance, by the parents. Those parents who choose for the analysed childcare investment have certain preferences and constraints that are likely to be different for parents whose children do not take part in the programme. Those differences in preferences and constraints might have resulted in a certain group of children being *selected in the treatment and* hence might have an impact on the child outcomes which might be misperceived as the actual effect of the programme. A great number of factors can influence the parents' choice for childcare arrangements, for instance, individual cultural tastes on childrearing, the incentives for parents to work, the availability and costs of non-parental childcare, and also the perceived quality of the childcare arrangement. Such parental selection based on unobserved and observed factors can influence both the choice of quality of childcare and the development of the child at the same time. Accordingly, the crucial identification problem is to first understand the assignment of the treatment and to find a strategy to account for potential selection

biases, unobserved endogenous differences and unobserved simultaneous effects that overlap and hence bias our estimated effects.

All individuals of the populations are usually not observed. Statistical inferences for the whole population evaluations of policies need to fall back on estimating Average Treatment Effects (*ATE*)[9] from summarizing what is known about an observed sample of the population. For the example of *ECCE* programmes, the average treatment effect can be defined as an expected Development Gain (*DG*) in the children's outcome variable which is caused by participating in the *ECCE* programme. This gain in the outcome would not be visible if the same child would not have participated. 'Expected' refers to the fact that there is a certain probability of being treated and therefore experiencing the Treatment effect on the Treated (*TT*), as opposed to the effect the treatment would have for those who are not treated (Treatment effect on the Untreated, *TU*). The problem is that in a (ex-post) policy evaluation *TU* – the counterfactual - is unobserved. A common assumption is that children who are not participating in the *ECCE* programme (control group) are actually comparable in their characteristics *X* to the children who are participating in the programme (treatment group), except the actual treatment of the *ECCE* investment, and that the characteristics *X* are unrelated to the participation in the *ECCE* programme[10]. In other words, to have consistent estimates it needs to be assumed that there are no confounding factors such as a non-random treatment assignment that might bias the estimates. Given this assumption, single differences between the mean child outcomes of the treatment group and the mean child outcomes of the control group could be used to get consistent and unbiased estimates of the *ATE*[11]. In doing impact evaluations of *ECCE* programmes, it needs to be addressed to which extent it can actually be assumed that programme participation is exogenous, conditional on all confounding

[9] $ATE(X) \equiv E(DG \mid X) = TT \cdot \Pr(X, ECCE = 1) + TU \Pr(X, ECCE = 0)$

$DG_i \equiv Y_i^{ECCE=1} - Y_i^{ECCE=0}$

$TT(X) \equiv E(DG \mid X, ECCE = 1)$

$TU \equiv E(DG \mid X, ECCE = 0)$

[10] This assumption is called 'Conditional Exogenous Assignment Assumption' or alternatively 'Conditional Independence Assumption (CIA)'.

[11] $DG(X) \equiv E(Y^{ECCE=1} \mid X, ECCE = 1) - E(Y^{ECCE=0} \mid X, ECCE = 0)$

factors. Statistical and econometric methods have greatly contributed over the last decades to improve identification strategies in the policy evaluation studies in economics. The growing attention of economists to the policy evaluation of *ECCE* policies, applying those identification strategies and economic concepts will contribute to a sophistication of the evaluation research done for *ECCE* programmes.

Outline

A number of identification strategies have been used in the literature to address estimation problems when evaluating social programmes. This paper is a contribution to the literature on evaluating treatment effects of social programmes[12]. The preferred strategy is the use of randomized experiments. Such experiments can take place in a laboratory setting, in the context of a social field trial or be documented by a randomized treatment assignment which results from a 'natural' exogenous event. Another strategy, involves the use of propensity score matching methods that can be used, for instance, when observations of an experiment are not sufficiently balanced. Following this, the most commonly used identification strategy is a multivariate approach which accounts for an extensive list of confounding factors. This includes a referral to value-added approaches, family fixed effects estimations as well as twin and adoption studies, this paper then elaborates on non-experimental approaches to identify causal effects of treatments, beginning with differences-in-differences applications and then expanding on the use of instrumental variable applications, including regression discontinuity designs. It continues with a discussion of quasi-structural estimation in which the counterfactual is explicitly modelled. Manski's non-parametric method of estimating bounds of the treatment effects is put forward as a straightforward test that does not demand extensive distributional assumptions. Finally, there is a discussion of the benefit of meta-studies in aggregating findings from a number of evaluations using the afore-mentioned identification strategies. The final conclusions consider how to choose the best identification strategy for evaluating treatment effects. The aim of this paper is to relate

[12] Valuable literature sources for identification strategies to evaluate social programmes are (Card, 1999; Heckman, 2008a; Heckman, Lalonde, & Smith, 1999; Heckman & Vytlacil, 2007; Imbens & Rubin, 2008; Manski, 1993; Ravallion, 2005; Todd & Wolpin, 2003).

those strategies to the evaluation of *ECCE* investments[13]. It contributes to the understanding of the importance of the application of identification strategies that are deemed to better be able to make causal rather than predictive inferences for evaluations of *ECCE* effects. This toolkit provides (new) researchers and people interested in the field of ECCE investments an introduction to the range of methodologies used for *ECCE* evaluations and a reference that can be used to find literature of applications of those methods.

[13] Another discussion of causal inferences in evaluations of early childhood programmes can be found in Shonkoff & Phillips (2000; chapter 4) and Currie (2003).

Identification Strategies

Randomized experiments

The best solution to ensure validity of the *conditional exogenous assignment assumption* is a randomized experiment (see, for instance, Rubin, 1974). In a randomized experiment on *ECCE* investments, children would be randomly allocated to either the treatment group or the control group, independent of their actual characteristics, in order to ensure that all characteristics are distributed equally in both treatment and control groups despite the difference in the allocated treatment. Randomized experiments can take three forms: 1) laboratory experiments, 2) social experiments, and 3) natural experiments. Laboratory experiments take place in a fully controlled environment, for example, at a lab in the research institution; the randomized allocation to treatment is usually determined ex-ante, before the experiment starts. Social experiments take place in the field and can be, for instance, pilots of social programmes, which are tested on a population sub-sample; in social experiments the randomized allocation also predominantly takes place ex-ante. The researcher still has some control over the allocation of the treatment; however, the environment in which the experiment takes place cannot fully be controlled[14]. In natural experiments, the researcher relies fully on an external source of random allocation of the treatment; the researcher is a mere observer. Hence instrumental information needs to be gathered ex-post which contains sufficient information on the random component of the allocation process. Natural experiments allow the researcher the least control over any confounding factors and restrict the availability of relevant data for evaluations.

Laboratory experiments are traditionally used by biologists and chemists. Lately they are also favoured by social scientists; some economists have been interested, for instance, in testing the validity of behavioural assumptions, such as the rational decision-making of people, which is the main basis for most of economic theory. In order to test and estimate theoretical models with, experiments should be kept as simple as possible in terms of treatments and should account for various characteristics that have an impact on the measured outcome variable. Laboratory research on child development is predominantly

[14] That is why such social experiments are also called 'partial experiments'.

conducted within mostly small-scale, well-observed groups in psychological research. Given ethical concerns, animals are often used to test learning theories as in Pavlov's famous dog experiment on learning through conditioning (Pavlov, 1927) or to test, for example, the attachment between mother and child as in the Harlow's experiments with infant monkeys (Harlow & Zimmermann, 1959). Despite the advantages of such animal research, the conclusions may not be valid for humans. This criticism is also valid if humans are participants of such laboratory experiments particularly because the type of persons joining such experiments is often not the same as type of persons in the population for which the theory is actually tested. It should also be noted that people who participate in laboratory experiments tend not to be typical of society as a whole – often students are the easiest available group to study, despite their limited representation of the population.

Nevertheless, laboratory experiments give the researcher the strongest locus of control over important factors. Laboratory experiments allow for strong internal validity in testing theories and models. They allow testing those even in ways that might not be observable in reality. In a experiment on *ECCE* investments, environmental factors such as the presence of family members, the pedagogic setup of the programme, the training levels of care staff, the doses of the programme in terms of hours per day or number of visits per week, or the group composition of peers can be much better controlled than in a non-experimental research design. However, such control often raises practical and moral concerns, which permit the use of such strict experimental setups. In this regard, it would be unethical to assign potentially harmful treatments, for example, low-quality *ECCE* investments, excessive exposure to TV, poor nutrition, pre-natal smoking, withdrawal of early affection, or loosening of attachment of close persons. The question is whether an *ECCE* investment should treat children like 'guinea pigs', when there are only theoretical hypotheses on its outcomes but no empirical evidence yet. Sometimes it is not yet clear which direction the effects might have, for example, whether high quality non-parental childcare is harmful or eventually beneficial for infants. Thus in such unclear cases ethical considerations might still leave some scope to conduct an experiment if it is unclear whether a treatment is likely to be harmful or not. To some degree, evidence from

other (non-experimental) evaluations of similar investments and case studies of best practices might tackle this problem of forecast-uncertainty of treatment effects.

Even if ethical considerations provide scope for doing an experiment, some treatment experiments might be impractical, for instance, sending children to external care rather than letting them stay with their parents, as parents would probably object such a decision (Shonkoff & Phillips, 2000). Some parents might decide that their children should drop out of the experiment or do not participate at all. The less incentive there is to participate the lower the burden will be of opting out in particular for those children who are not in need of the experimental treatment. This could bias the results from the experiment.

Moreover, the strong control in particular in laboratory settings increases the probability that recruitment procedures to receive treatment could be substantially different from the recruitment of a concrete programme, putting doubts on external validity. The number of factors that are used to account for equal distributions in treatment and control group is always limited by the data available. Consequently a different group might be exposed to the randomized treatment than the group that is joining the real programme later on, resulting in a 'randomization bias' (Heckman & Smith, 1995). The strong control also raises doubts of external validity. Real life conditions might be too different from experimental conditions - making inferences from the experiment participants to the whole population invalid. The might still be the case even if conditions in the experiment are adapted as much as possible to reality and if bigger sample sizes are used. The stronger the environment is controlled and, accordingly, the more people are aware that they take part in experiments the more likely participants might be to behave unnaturally or atypically; this so-called 'Hawthorne effect' might bias our conclusions (Landsberger, 1958). There might also be a problem of endogenous groups making different behavioural choices, for example, some might plan to stay on in a programme once the experiment is finished while others do not have this ambition. Often experiments are not incentive-compatible with reality, in monetary terms but also in terms of trust and expectations (Glaeser, Laibson, Scheinkman, & Soutter, 1999). What a participant states during an experiment might not necessarily be true and instead be mere 'cheap talk'.

Accounting for such incentives structures can hence improve external validity of estimation results. Non-experimental supplemental evaluations similarly improve the credibility of external validity (Moffitt, 2004).

Social experiments are easier achievable as they demand less strict controls (Heckman & Smith, 1995). However, the political feasibility is often a serious concern. Randomizing the treatment will make some of the children participating in an *ECCE* programme who are actually less in need of such a programme than a number of children who are actually more in need but have been randomly selected to the control group instead. If resources are limited, targeting them to the most in need might be a practical priority. Usually, policymakers want then to aim the *ECCE* programmes at those children who are most in need, who have certain observed characteristics, for instance, documented delays in development, or an 'at-risk' socio-cultural environment associated with an increased likelihood of development problems or, children whose family structures or parental employment patterns potentially result in less time available parental care. When social experiments are targeted, it is more likely that biases from potential macro feedback effects are neglected (Garfinkel, Manski, & Michalopoulos, 1992).

What would the effects be if the *ECCE* experiment goes to scale? For instance, whereas a well planned and controlled small scale programme, such as the Perry Preschool programme, can achieve big successes, it is much more difficult for larger scale programmes (e.g. the Head Start programme) to achieve the same success, as they are, by their very nature, less successful in targeting precisely and in replicating treatment environments of small-scaled programmes (Ramey & Ramey, 2003). Clear reference to the partiality of the analysis as opposed to the general context is recommended. Social experiments are also less able to control over spill-over effects between participants, for example, that parents from the treatment group exchange views with parents of the control group and vice versa, bringing about contamination problems. Further serious threats to the validity of social experiments are that local media might pick up the topic or that policymakers deliberately give political statements which influence expectations and, hence, the preferences and behaviour of participants in the experiment. Bad planning of

an experiment might not only allow such spill-over effects but even demoralise the stakeholders involved in a social experiment, ridiculing the continuation of heavy investment of resources in the long-term (Shonkoff & Phillips, 2000).

Social experiments are still rarely conducted into *ECCE* investments as they need a lot of careful preparation and – similar to laboratory experiments - raise doubts regarding their external validity. However the general tendency towards evidence-based policymaking, particularly in the education sector, has increased the willingness of actors in the *ECCE* arena to endorse social field experiments. The predominant examples of randomized social experiments referred to in the literature are the US American Perry Preschool (High/Scope) programme in Ypsilanti, Michigan and the Abecedarian programme in Chapel Hill, North Carolina[15]. The Perry Preschool Programme was a half-day preschool programme that was randomly assigned to a group of about 65 African American 4-6 year old children of a sample of 123 children from 100 families from 1962-1967. The treatment group followed a two year programme of about 2.5 hour classes per day. The children of both the treatment and a control group were monitored until the age of 40. Whereas the control group children received no pre-school. All children were living in poverty. The Abecedarian programme was a full-day programme for children aged from four months to five years, born between 1972 and 1977 that were selected randomly from a sample of 111 children in 109 families. Follow up observation was carried out on the randomized treatment and control group until the age of 21 years. Evaluations of both programmes accounted for a number of observables, in particular related to the parents, and both revealed that such *ECCE* programmes can have significant positive long-term outcomes such as increased achievement test scores, lower propensities for grade repetition, time in special education, criminal and delinquent behaviour as well as a higher propensity to graduate high school. While these positive outcomes are very promising, they need to be understood as occurring at the extreme cases of high quality

[15] See, for instance, (Campbell, Helms, Sparling, & Ramey, 1998; Cunha, Heckman, Lochner, & Masterov, 2005; Schweinhart, Barnes, & Weikart, 1993). An interesting social experiment is also the Nurse Home Visitation Program which assessed the impact of randomly assigned home visits by nurses to mothers in semi rural areas of New York in the period 1978-1980 for 15 years (Olds et al., 1998). Another often cited example is the STAR project in Tennessee which is a social experiment that took place during the post-kindergarten period in the time of 1985-86, testing the impact of class-size and student-teacher ratio on randomly assigned children who entered primary education (Krueger, 1999).

and strictly targeted treatments. Hence, such social experiments might be relevant for testing a theoretical causal relationship but they are still limited in providing fitting answers for the all children who benefit from such programmes.

When such laboratory or social field experiments are not feasible for practical or ethical reasons, researchers might like to use an array of alternative strategies to tackle biases in results of estimations which are based on non-random assignments. Such strategies are also called natural or quasi-experiments; the researcher tries to find an exogenous element of randomness in the assignment process that is not related to the outcomes observed. A number of sources might provide such elements, for example, sudden changes in programme rules, discontinuous eligibility criteria and access limitations, natural causes of randomness and supplementary choice information. Those sources and the related identification strategies will be discussed later in this paper.

Propensity score matching

Even if it is assumed that the treatment assignment is exogenously given with respect to individual characteristics, as given in the afore-mentioned random experiments, attrition and non-response might still result in unbalanced treatment and control groups. Unbalance refers to the sample not representing the whole population anymore. If this unbalance is due to observable factors, a sufficiently large overlap of characteristics in the two groups might be found; matching overlapping sub-groups could re-establish a balance. Matching demands good literature understanding of potentially influential characteristics. The more information is available in terms of the unit characteristics in both groups, the more precise our matching will be. Newer datasets are richer because they cover a large range of characteristics and larger samples. These datasets allow more sophisticated accounting for confounding factors and, hence, provide a prospective option for better matching. Nevertheless, doubts often remain about accounting sufficiently for unobservable influences on assignment such as pre-programme outcome levels that ridicule the exogeneity assumption (see, for instance, Agodini & Dynarski, 2004). Despite such likely neglect of relevant factors, matching studies are also sensitive to measurement errors with respect to the characteristics that are included in the matching.

Different matching techniques have been developed for this exercise that are free of the ethical and practical considerations connected to randomized experimental setups once sufficient data has been collected ex-post. As average treatment effects are assessed, a one-to-one match of children with exactly the same characteristics is not necessarily needed - despite of course their difference in *ECCE* treatment. When the distributions of the characteristics are overall the same - are in balance - causal inferences can be drawn without attrition or response biases. Matching can be done exactly using categories of one or more variables. However, if the number of variables and dimensions of characteristics increase, the algorithms to match according to categories get increasingly complicated. Moreover, the more dimensions are accounted for, the more difficult will it be to find exact matches. Propensity Score Matching (PSM), as proposed by Rubin and Rosenbaum (1984), is an estimation of the probability of participating in the programme along a

larger number of observed characteristics and thus overcoming possible limitations of exact matches. It allows for matching individuals of the comparison and treatment group[16] who have similar propensities for certain characteristics. The simplest PSM method is pair-wise (1-to-1) matching in which one observation of the treatment group is matched by one observation of the comparison group which has the same propensity scores indicating their similarity (Imbens & Rubin, 2008: sect. 14.2 Exact Matching Without Replacement). Other alternatives are 'nearest-neighbour matching' in which each observation is weighted by the neighbourhood distance and 'calliper matching' in which the nearest neighbour is found within a certain range (Imbens & Rubin, 2008: sect. 14.4 Distance Measures). Multiple (1-to-n) matches are also possible (Imbens & Rubin, 2008: Section 14.3 Inexact Matching Without Replacement); they have the advantage of lowering the variance in the estimators (Dehejia & Wahba, 2002). Kernel matching estimators are an example of multiple matches. In this case the whole comparison group is used to measure individual weights which are covered in a function that catches the distribution of matched characteristics (Bryson, Dorsett, & Purdon, 2002).

The propensities to receive an *ECCE* treatment are unknown to the researcher. Hence, as a first step it needs to be understood how participation in the treatment is related to observed characteristics of children and their parents, in case those characteristics are determinants of the choice to participate. Accordingly, it needs to be understood whether the child benefits from a single treatment (e.g. a parent-child programme, or multiple *ECCE* investments (e.g. a parent-child and a preschool programme). Therefore it needs to be ensured that only those variables are included which influence both participation and outcomes in a choice model - a binary choice model can be used for a single investment, a multinomial choice model for multiple investments. Heckman, LaLonde and Smith (1999) point out that no variables should be included that might be affected by the *ECCE* investment itself or by anticipating it (e.g. labour supply of mothers), so as to prevent adding new biases. The overlap in characteristics of both members of the treatment group and the control group is often called 'common support'. The number of variables

[16] Usually matching searches for units of comparison for observations of the treatment group, hence the term 'comparison group' is often used in matching studies rather than 'control group'.

included in the PSM process should also be in balance with the size of the common support. Each additional variable exponentially increases the multivariate dimensions necessary to find a common support area.

As PSM brings along additional variation beyond the usual sampling variation, the standard errors of the estimates need to be corrected (Heckman, Ichimura, Smith, & Todd, 1998). Bootstrapping can be used to account for this problem as long as the sample size is sufficiently large. In applying this method, random draws from the given sample are taken, with a large number of repetitions so as to create a large number of randomly reordered datasets[17]. In this regard, bootstrapping relies on the statistical distribution as it is in the data, it does not have to make any parametric assumptions. A too small common support can give some indication of selection bias due to excluded factors which are non-overlapping (Bryson, Dorsett, & Purdon, 2002). If the overlap is too small, then PSM might not be the most accurate method to use as the external validity of estimation results on the overlapping sub-sample might be limited.

The advantage of PSM as compared to the usual accounting for characteristics in (linear) regression function lies in the non-parametric approach. It allows for more robust predictions within the area of common support as it does not make any extrapolations outside of that area (see e.g. Cochran, 1963). Another advantage of PSM is that it gives the possibility of addressing heterogeneous treatments by doing matching within certain sub-groups that have the same treatment. A drawback might be that PSM methods do not allow a look at the distributional effects of programmes, which is often a policy relevant question for example when looking at the potential social mobilization of disadvantaged children through *ECCE* investments. The inability to recover mean (homogenous) effects is a joint limitation of PSM and other techniques such as experimental designs (Heckman, Smith, & Clements, 1997).

[17] The special version of 'jackknife' bootstrapping is based on dropping observations at random in the bootstrapping process.

There has been relatively little use of PSM in the evaluation literature of *ECCE* investments and programmes so far. The core reason is that the researcher cannot realistically assume that treatment is exogenously given conditional on unobservables even if newer studies are used, with large samples and extensive information on observable pre-treatment characteristics. Two US studies that attempted a PSM method to evaluate effects of different preschool arrangements are those of Henry et al. (2006) and Loeb et al. (2005). They attempt using a maximum of information on pre-treatment characteristics such as early pre-treatment assessments. To test the validity of the PSM estimates, the second study even used childcare supply as an instrumental variable for the preschool treatment. Despite those efforts, they still cannot claim to have reached full equivalence in matching with respect to exogeneity of treatment and hence to have unbiased estimates. Behrman, Cheng and Todd (2004) combined PSM methods with estimating marginal effects of Bolivian preschool programmes. They make the strong assumption that unobserved selection would be independent of the dose of treatment (duration in the preschool programme) and, therefore, that unobserved choice factors would be exogenously given in studying marginal effects using PSM methods. Some other evaluation studies of early child development programmes in low and middle income countries combine PSM with a difference-in-difference approach[18] to correct for unobserved biases through accounting for average earlier outcome observations (see e.g. Armecin et al., 2006; Gultiano, 2006).

[18] The difference-in-difference approach is explained later in this paper.

Multivariate approach

The majority of studies evaluating *ECCE* effects rely on coping with selection biases by accounting for confounding factors through inclusion of rich arrays of control variables in a non-experimental ex-post research design. A number of confounding factors might moderate the treatment effect. Shonkoff and Phillips (2000) mention the example of maternal depression, which might be related to a lower participation in home visit programmes. Accounting for depression would allow for differentiating the effect of the treatment from the effect of less responsiveness of depressed mothers to the treatment. Controls can include a vast array of variables on parental characteristics such as parental marital status, incomes, occupation, education and ethnicity, on children's characteristics such as sex, age, pre-treatment development status and number of siblings as well as control variables for other *ECCE* investment modes and variations in qualities. Controls can also include latent variables such as parenting style, socio-economic status and childcare quality indicators such as environmental rating scales (e.g. for infants and toddlers, ITERS; for early childhood, ECERS; for family childcare, FCCERS; and for children of school age, SACERS[19]).

The common conviction of the multivariate approach is that the more information on confounding factors is included, the less scope for a bias in the estimated *ECCE* effects is left. Accordingly, the richer the dataset used is, the more precision is expected of those estimation results. Innate to this approach is the exogeneity assumption of the treatment conditional on any characteristics. By implicitly saying that once the numerous controls are included to correct for potential selection biases, it can be assumed that there is no endogenous treatment allocation left. A number of researchers that follow this approach seem to stress that if you start putting in more and more controls and if you do not get any additional effects on the coefficient of the *ECCE* effect anymore that you then could be sure that there is not much more selection bias left (see e.g. Ruhm, 2000). This logic has an important flawed assumption, though. It assumes that the controls can pick up - through direct or latent effects - all the potential influences on the child's outcome that

[19] For more information on those rating scales refer to the following website: http://www.fpg.unc.edu/~ECERS/index_frames.htm.

might otherwise mistakenly be interpreted as a part of the treatment effect. It is rather likely that key unobserved factors are still left out and hence selection bias is never perfectly accounted for. Even with the richest dataset on child development it might not be possible to measure all relevant aspects; information, for example, on confidence, motivation, genetic disposition and general abilities might still be limited. The large variation in estimation results in the *ECCE* evaluation literature due to application of different multivariate specifications is an indication that this method is not fully successful in accounting for selection biases.

Moreover, this approach assumes that if characteristics that are used as control variables would actually be different in the counterfactual case; behaviour would change according to the expected sign of the coefficient of the control variables. But this is not a valid check of the counterfactual case as long as participants might not change their behaviour due to any other unobserved variable which is not accounted for. Given the problem in properly accounting for the counterfactual leaves a lot of scope for spuriousness and hence limits the power in drawing causal inferences from the estimates. For this reason, such studies should be sufficiently agnostic with respect to interpretations of estimation results; these allow predictive rather than causal inferences. This clear differentiation is often not made explicitly, though.

The multivariate approach is the most commonly used approach, in particular within sociological and child development research, to evaluate the treatment effects of various early childhood care and education investments[20]. The US American NICHD-ECCRN[21] study and the UK EPPE (3-11)[22] studies provide several examples of evaluations using multivariate approaches to evaluate various *ECCE* investments (e.g. Sammons et al., 2002, 2003; Sylva et al., 2003). Subsequent to a profound descriptive study of the data,

[20] (see e.g. the evaluation studies of Baydar & Brooks-Gunn, 1991; Buechner & Spiess, 2007; Leibowitz, 1974; K. Magnuson & Waldfogel, 2005; K. A. Magnuson, Meyers, Ruhm, & Waldfogel, 2004; Vandell & Ramanan, 1992)

[21] Abbreviations for the National Institute of Child Health and Human Development (NICHD) and its Early Child Care Research Network (ECCRN). For more information see: http://secc.rti.org/.

[22] Abbreviations for Effective Provision of Pre-School Education (EPPE) and Effective Pre-School and Primary Education 3-11 (EPPE 3-11). More information at: http://www.ioe.ac.uk/schools/ecpe/eppe

this approach generally provides a straightforward access to understanding associations between relevant factors of early childhood and their relation to the child outcome variable(s) of interest. The inclusion of numerous explanatory factors in the estimation provides potentially rich predictions of several effects. However, (initial) strong assumptions of externality of treatment allocation process need to be made. Unfortunately, those assumptions are not always put forward in the multivariate evaluation literature.

Value-added approach (child fixed effects)

Initial human capital endowments are important to understand differences between children. Multivariate estimations often include several proxies for such endowments, as for example, parental human capital indicators and earlier development outcomes of children, as proxies for potential genetic endowments as well parental investments. However, such proxies are only second-hand solutions to account for endowments. The growing availability of longitudinal data on children before and after treatment makes up for some of the weaknesses of this research strategy by accounting sufficiently for pre-treatment differences in children and their parents with regard to possible choice behaviour and non-treatment related stimuli for the children. In such value-added approaches lagged test scores and other indicators of ex-ante endowments are used as control variables (see e.g. Duncan, 2003; J. F. Ermisch & Francesconi, 2000; Hsin, 2007). However, applications of such valued-added models in *ECCE* evaluations are often constrained by the availability of good pre-treatment outcome information. This is particularly true if the available dataset does not have sufficient longitudinal observations. It is also debated whether the development status of children before entering a preschool programme, for example, at the age of three can in fact be properly tested. Other critical questions are: Do these test outcomes reflect all important factors of the child's ex-ante development status? And do these tests sufficiently overlap with later outcomes in the abilities that are assessed and those which are supposedly influenced by the *ECCE* investment? Despite the need to understand the various test outcomes that are used in *ECCE* evaluations, applications of value-added approaches demand a good understanding of the development domains that are assessed.

Family fixed effects

Another extension of multivariate estimations is the use of Family Fixed Effects (*FFE*) models (= sibling differences) as done, for instance, by Waldfogel, Han and Brooks-Gunn (2002). They estimate effects of post-natal maternal employment on children under three. First they added controls to catch the omitted variable bias of observed factors such as the home environment, the use of breast-feeding and the attended type of childcare. Additionally they used family fixed effects to account for unobserved heterogeneity between mothers. Mothers who are less interested and skilled in providing care and stimulation for the child might be more likely to return to work early. A limitation of *FFE* is that it covers only characteristics that are fixed over time (and across siblings). If mothers change their employment status or adapt their parenting preferences for the next child, the *FFE* estimations might still be biased.

In a *FFE* approach, it is assumed that children of the same family (or in a broader version, the same household) have been influenced by the same unobserved factors such as parenting styles or parental preferences in choosing certain *ECCE* investments. In some cases family effects are fixed not across siblings but across ages of the same child, assuming that family influences are not changing across ages. In the case of *FFE* across siblings, it is assumed that those factors are fixed across the children. It implies that, for example, parents do not change their parenting style across their children or adjust their preferences for an *ECCE* investment from one child to a later one (Todd & Wolpin, 2004). There are several application within *ECCE* research that benefit from large datasets that allow family fixed effect estimations (see e.g. Janet Currie & Thomas, 1995; James-Burdumy, 2005; Lefebvre & Merrigan, 2000; Rosenzweig & Wolpin, 1995; Ruhm, 2000). *FFE* estimations account for numerous nurturing influences of the family that are stable across children. As long as *FFE* are taken across siblings of the same biological parents, such *FFE* also account for some genetic influences that partially determine the child's development patterns. As mentioned earlier, such endowment effects can also be accounted for by using pre-treatment development indicators.

Twin studies

Genetic endowments vary between siblings or pre-treatment outcomes contain measurement errors. Usually social scientists are interested in studying the true causal effects of nurturing factors in order to propose the best *ECCE* programmes. The best account for unobserved genetic endowments is the study of monozygotic twins as they are equal in terms of their genes and experienced the same pre-natal conditions. Still, differences in birth weights and moments might result in differences in endowments. As the natural likelihood of having (monozygotic) twins is rather small, huge datasets are necessary to get a sufficiently large sample size to use twin fixed effects estimation. And if a sufficiently large sample is found, issues of censoring and sample bias need to be addressed to guarantee external validity of the estimation results. Even if there are more observations of siblings in a dataset, there might still be a limitation in finding enough sibling pairs where siblings differ in terms of their treatment.

Adoption studies

Another extension of family fixed effects models are adoption studies which can be used to determine the importance of environmental influences as children of similar genetic endowments are brought up in different environments (see e.g. Björklund, Lindahl, & Plug, 2004; Plug & Vijverberg, 2001a; Plug & Vijverberg, 2001b). Such adoption studies make a few assumptions to guarantee internal and external validity (Holmlund, Lindahl, & Plug, 2008). They assume 1) that children are randomly assigned to the adoptive families; 2) that adoption takes place at birth, 3) that children do not diverge in their development due to differences in their adoption experience; 4) that adoptive parents do not differ from other parents; 5) that genetic endowments passed from biological parents on to their children are not related to their resources and their decision to give their child for adoption.

The potential weakness of multivariate approaches in identifying causal effects is their reliance on observable factors accounting sufficiently for omitted variable biases. Application of fixed effects models can only result in consistent estimates if those fixed effects cover the full scope of potential unobservable characteristics.

Differences-In-Differences Approach

Another strategy that is often used in accounting for unobserved heterogeneity of treatment and control groups is the difference-in-difference (*DID*) approach (see e.g. Blundell & Macurdy, 1999). The *DID*-approach can be seen as a relaxation of the value-added model to comparisons of pre- and post-treatment outcomes of potentially different individuals. The innate concept of this approach is rather simple. Inferences are made through correcting for ex-ante differences in outcomes in a first step, assuming that in this way all unobserved factors that have influenced the outcomes in the first place have been accounted for, and then in the second step taking another ex-post difference for the changes in outcomes that resulted from the treatment. In the context of *ECCE* evaluation research, pre-treatment child outcomes are used to account, for instance, for unobserved differences in abilities (genetic and nurtured), parenting styles and parental choice behaviours. Despite accounting for unobserved differences through the first difference, *DID*-applications are usually combined with inclusion of multivariate controls so as to account also for observable differences.

The identifying assumption is that the ex-post outcomes of the comparison group and the ex-ante outcomes average out to zero across all the individuals who participated in the *ECCE* programme. By taking averages across individuals, repeated cross-sections can be used – longitudinal observations of the same individuals are essentially not needed. However, it must be stressed that the repeated cross-sections have the same compositions of individuals so that averaging across individuals according to the identifying assumption can indeed be zero. Another essential assumption underlying this approach is that those unobserved effects do not change during the treatment period, for instance, that parents don't change their parenting style while the child is attending an evaluated preschool programme or that the macroeconomic situation and lifecycle situation of parents doesn't change. Individual effects are fixed over time, and transitory trend effects between repeated observations are assumed not to be related to participation in the programme – in other words, treatment and comparison group are affected by such changes in the same way.

This assumption is, however, not very realistic. If there is serial correlation, the estimates might grossly under-state the standard errors around the estimated treatment effect (Bertrand, Duflo, & Mullainathan, 2002). In some cases some information might be available on the changing patterns of some of the factors that are accounted for in the first difference. A difference-in-difference-in-difference (*DIDID*) approach could be used if there is concern that the estimated treatment effects of a *DID*-application are spurious. The regression-adjusted outcomes might show different trends between treatment and comparison group. If this is the case, a third difference might be used to account for those changes. Also, a comparison group might be found that is supposedly similar to the treated group, but which has actually not been treated. However, such a third difference can only be based on observed information, it cannot account for changes in unobserved factors. One of those unobserved trends might be related to what is called an Ashenfelter's dip. Ashenfelter showed that a temporary reduction in earnings just before a training programme resulted in a higher participation in the programme. Without accounting for such unobserved 'before ex-ante' factors, the *DID*-estimators of training effects might be overestimated (see e.g. Ashenfelter, 1978; Heckman & Smith, 1999).

Instrumental variables

The use of instrumental variables (*IV*s) as non-experimental empirical strategy is a preferred research strategy as it tries to make use of exogenous sources of randomness that might make it possible to replicate an experiment and hence overcome potential selection biases due to unobserved factors (Angrist & Krueger, 2001). An instrumental variable is a variable that contains some information on the treatment assignment but is unrelated to the outcome. In a two-stage estimation approach, the information on the treatment is first extracted from the instrumental variable(s) by running a regression (Theil, 1953). The predicted values contain information on the treatment which is only related to the instrumental variable(s). If the *IV*s really are uncorrelated to the outcome variable, the predicted value can be used in the second stage as an instrument for the treatment when estimating the treatment effects. Usually researchers face difficulties finding good instruments as they need strong arguments to stress the independence of the instrumental variable and the outcome variable and correlation between the *IV* and the endogenous variable. The search for an *IV* is guided by the researcher's speculation. To get consistent estimates there need to be at least as many instrumental variables available as endogenous repressors. In *ECCE* evaluation research, the use of instrumental variables is still in its infancy as arguments on the *IV*s independence of the child outcome are often weak or as the *IV*s do not identify plausible effects of the *ECCE* treatment on the child outcome.

A number of diagnostic test such as the Anderson-Rubin test, Hansen's j-statistic, a first stage F-statistic and the *IV*'s significance in the first and second stage of the estimation allows us to assess our assumptions (for a discussion of those tests, see Stock, Wright, & Yogo, 2002). Monte Carlo simulations (MCS) can be used to test the properties of our estimates. In this parametric method, the whole model is simulated (including the *IV*s), with a large number of repetitions using different sets of random numbers as inputs. The outputs generate by MCS provide ranges annotated by probabilities rather than fixed value for the estimated treatment effect. With the MCS on the *IV* estimations can benefit from comparing the density functions of the *IV*s with the density functions of the OLS

biases. Hausman tests might be used to test for the correct choice of a subset of instrumental variables of a wider set of potential *IV*s (Hausman, 1978).

When making use of an *IV* strategy strong implicit assumption are made about homogeneity of the treatment effect or at least that all relevant differences that make people react differently to the treatment are accounted for. In this case the local average treatment effects (*LATE*) (Imbens & Angrist, 1994), the effect of the treatment on the treated (*TT*) (Heckman & Robb, 1985) and the treatment effect on randomly selected observations from the population are the same. If treatment effects are heterogeneous, additional assumptions need to be made on how people act according to the *IV*'s information with respect to their participation in the treatment (Heckman, 1997). For instance, in *LATE* estimations, it needs to be assumed that people do change from non-participation to participation in the treatment if the instrument for which *LATE* are calculated is changed.

Candidates for *IV*s have been, for instance, variables informing about the availability of childcare or any alternatives for estimating the effect of childcare or preschool attendance. In this case *IV*s could be the variation in the availability of family members and friends, the local public expenditures on *ECCE*, eligibility criteria, guaranteed places and waiting lists for *ECCE* arrangements and information on childcare supply (for applications see e.g. R. Bernal & M. P. Keane, 2006; Würtz, 2007b). To estimate the effects of maternal employment on infants and small children, researchers tried using instrumental information, for example, on employment and childcare regulations, labour market conditions, maternal leave policies, availability of child benefits and care subsidies or previous experience in combining maternal employment with childcare time (see e.g. Bernal, 2007; Blau & Grossberg, 1992; James-Burdumy, 2005; Würtz, 2007a). Other studies used sudden (exogenous) shocks to the family as instruments for *ECCE* choices; for example, income losses due to (sudden) parental disability, lottery wins, unanticipated macroeconomic shocks and other severe events. In this regard, policy reforms affecting *ECCE* choices provide potentially good instrumental information as long as it can be argued that those reforms are exogenously made and not influenced by the parents.

Regression Discontinuity (RD) Design

Such policy reforms often lead to a discontinuity in the treatment. At a certain cut-off date (or point) treatment is changing. This discontinuity can be used to estimate treatment effects by comparing the groups before and after (below and above) the cut-off date (point). In some cases eligibility criteria that are unanticipated by the individual or geographical boundaries can provide also cut-off points that force different participation in the treatment. Sharp cut-off points stress a 100% change in treatment while 'fuzzy' cut-off points force changes in the dose of treatment. Around cut-off points, observations need to be chosen so that access to treatment changed while the ceteribus paribus condition holds - all other factors stay constant (see a discussion of RD designs, see, for instance, Imbens & Wooldridge, 2007). A graphical presentation of the regression discontinuity enhances the credibility in the choice of the best range around the cut-off point. The narrower the chosen range is, the better the fulfilment of the ceteribus paribus condition. The wider the range is the bigger the sample and, accordingly, the bigger the external validity of the estimation. A number of evaluation studies of *ECCE* investments has made use of such regression discontinuity designs (see e.g. Dustmann & Schönberg, 2008; Ludwig & Miller, 2006; Wong, Cook, Barnett, & Jung, 2008). The focus on a restricted range around a cut-off point often limits the validity of the results in making inferences for the whole population. Moreover, it might take some time before people adapt to a reform. This can be for many reasons, e.g. a lack of information. Without knowing those reasons explicitly it is very likely that the chosen range includes confounding cases, resulting in biased estimates.

Quasi-structural estimation

While an exogenous factor in the *IV* approach is used to account for participation choice for the evaluated treatment, quasi-structural estimations require specifying a functional model of this choice[23]. The counterfactual case is explicitly constructed by using factors which can be observed and are related to the choices made. As it is a structural approach, it explicitly uses theory in addressing selection biases and also allows heterogeneous treatment effects to be addressed – *IV* estimates demand extra assumptions or moment conditions to address such heterogeneous treatment effects (see e.g. Heckman, Urzua, & Vytlacil, 2006). Using such an approach for evaluations of *ECCE* investments would imply first modelling how parents choose, for example, their labour supply and the childcare arrangements and then estimating this jointly with the child's production function. The production function is a functional representation of the relationship between the child outcome and the explanatory factors, as for example, the *ECCE* treatment and the characteristics of the child and family. The unobserved heterogeneity in the child outcomes and treatment variable that is related to the participation decision can be accounted for by the so-called 'Heckman Selection Estimator' which is a latent variable evolving from the inverse Mills ratio of the participation decision (Heckman, 1976, 1979). If the observed information on factors related to the decision making is limited, this estimator is a Limited Information Maximum Likelihood (*LIML*) estimator; iterative methods provide the likelihood solution. Full Information Maximum Likelihood estimators (*FIML*) are statistically preferable, but computationally more complex as well.

As with the IV approach, good instruments to be included in the selection model need to be found first. As the aim is to estimate the participation in an *ECCE* programme in a first step, a prediction of the outcome is not desired. The factors included in the participation estimation of the first step should not be correlated with the factors used in the outcome regression (and hence be unrelated to the outcome). Consequently, the application of such a Heckman selection approach should begin with the identification of sufficient participation determinants that do not have correlation problems with the outcome

[23] This participation choice model is often called 'control function' as it controls for selection biases.

function (Little & Rubin, 1987; Puhani, 2000). This implies that the quasi-structural estimation relies on strong distributional assumptions relating to the unobserved variables. It should be assumed that the errors in the outcome equations and the participation model follow bivariate normal distributions (Heckman, 1979, 2008a)[24]. Monte Carlo simulations are often used to assess the strengths of the Heckman selection estimator (Puhani, 2000). As the covariance matrix evolving from the second stage regression is inconsistent, correct standard errors (and other statistics) should be generated, for example, through bootstrapping.

When assessing the causal effects of *ECCE* investments the decisions taken by parents regarding the child's participation are central – participation decisions taken by the young children and or public choices regarding provision of opportunities to parents and children are usually neglected (Haveman & Wolfe, 1995). Parents' decisions regarding *ECCE* investments are constrained by a number of factors, for example, public welfare and leave provisions including childcare services, the family structure, the availability of relatives and close friends who could care for the child, access to information as well as the financial, cultural and social capital that may be used to invest in early childhood care and education. Parental choices for *ECCE* investments are determined by their own preferences for leisure and consumption as well as for nurturing and raising their child which, in turn, are influenced by their perceptions of their children's needs. These factors show how complex it is to model parental decision-making for *ECCE* investments (see e.g. Leibowitz, 1974; Figure 1). The complexity of such multi-stage structural models in terms of various lagged factors and the joint determination of some factors implies serious modelling complications. The estimates of complex models are difficult to interpret. Above all, the data provisions on all the necessary factors for determining such models are also often incomplete.

[24] The recent trend to relax distributional assumptions in a number of empirical applications through use of non-parametric approaches is valid particularly for structural models. A number of non-parametric solutions to estimate treatment effects have been suggested. An example is the estimation of non-parametric bounds of the treatment effect as proposed by Manski (1989; Manski, 1990); explained below. See also Heckman's discussion of non-parametric evaluation methods (1990; Heckman & Vytlacil, 2000).

Recent applications of such structural models in *ECCE* evaluation are, for instance, done by Bernal and Keane (2006), Bernal (2007) and Würtz (2007a). Bernal and Keane estimated single mothers' employment and child-care decisions jointly with the production function of child cognitive ability. Bernal estimates a structural model on married mother's decision-making regarding work and day care, and how these affect child cognitive outcomes. Würtz models the parental decisions to invest their time in childcare, looking at households with two parents and one child. Ermisch and Francesconi (2000) develop a static model in which parental preferences are separable in their own consumption and their children's well-being. Thus they account for parental choices when estimating parental employment during childhood, and then the subsequently the educational outcomes of the children. A static model implies seeing childhood as a single period, dynamic models account explicitly for dynamic effects across several observation periods. In the context of *ECCE* investments dynamics can also be considered by splitting the early childhood up into a number of sub-periods such as pre-natal, infant, toddler, preschool and primary school period. Dynamic models look at the interaction of inputs and outcomes of earlier sub-periods on the outcomes of later sub-periods.

Manski bounds

The evaluation literature shows a trend for using more non-parametric methods in evaluating treatment effects while using fewer underlying assumptions on covariate distributions. One of those non-parametric methods, the bounding of treatment effects, has been suggested by Charles Manski. He suggested that even if no prior information is available on the selection process bounding can help to identify the range in which the treatment effect lies (Manski, 1989, 1990). *"Identification is not an all-or-nothing proposition. Although we may not have rich enough prior information and data to infer the exact value of a parameter, we may nevertheless be able to bound the parameter"* (Manski, 1993). The more concrete the available information on the covariate distribution of the outcome and respectively the treatment variable with other characteristics is, the closer can this range be drawn. Censoring the treatment and outcome variable to a certain possible range provides already minimal information for the creation of bounds of the treatment effect. If full prior information is available the bounds overlap on the point estimate of the treatment effect. Given the prior information on factors that might confound the outcome, the selection probabilities and then the bounds of treatment effects - conditional on those confounding factors - can be routinely estimated.

Application of Manski bounds in *ECCE* evaluations are done, for instance, by Ermisch et al. (2004) and Francesconi et al. (2005). Ermisch et al. estimate the effects of family structures (in particular single-parenthood) and parental joblessness during childhood on a number of child outcomes in early adulthood using British panel data. Besides other estimation strategies such as family fixed effects, they also calculate bounds along 144 combinations of confounding factors such as gender, age, birth order, parental education and mother's age at birth. Bootstrapping is used to calculate standard errors. Francesconi et al. do a similar analysis based on German panel data and calculate 96 combinations of confounding factors. Both analyses indicate that the precision of the bounds depends on the number of different groups and whether the within-group sample number (of each combination) is sufficiently large. Moreover, bounds of both analyses include zero which makes such bounds rather uninformative regarding drawing conclusions about the

treatment effect. An additional concern, not raised by those studies is the weakness of this approach as to measurement errors in the characteristics that are used for bounds of the effect. The higher the measurement error, the less correct are the calculated Manski bounds.

Meta Analysis

There is an enormous variety of conclusions drawn from evaluations of various *ECCE* investments using one or several of the afore-mentioned research strategies. Each of those strategies has its own strength and weaknesses and empirical limitations. Given the variety, the research field of *ECCE* evaluations is in need of comprehensive meta-analyses. A meta-analysis is a strategy to draw inferences on treatment effects from secondary sources by providing an inventory of findings. Such an inventory is done with careful reference to the strengths of chosen identification and empirical strategies. Such inventories can take often the form of a tabled overview. If such an inventory shows strong agreement on a specific treatment effect across different applications of research strategies as well as different population contexts, there is more assurance of having generally consistent causal inferences of this specific treatment (Shonkoff & Phillips, 2000; chapter 4). The use of meta-analyses is popular particularly in psychological and medical studies (Cooper & Hedges, 1994), for example, to provide inventories on the development outcomes of numerous clinical studies with children and animals. Hanushek has introduced the approach of meta-analysis as identification strategy to economics. He suggested using the estimates of each evaluation study as an observation for another empirical study – doing a regression on the regression coefficients (Hanushek, 1974; , 1977). Application of this Hanushek's concept to the evaluation of *ECCE* investments would imply that a large number of studies of a comparable *ECCE* investment, with comparable child outcome measurements in comparable contexts would be available. So far such comparable figures are not yet available. Once there is a convergence in *ECCE* investments, in measurements of child outcomes as well as in the data collection on context variables, meta-analyses might become an option. Despite the advantage of summarizing numerous contributions in an appealing overview, a drawback of meta-analyses is that sources of bias are not accounted for. A major source of bias is the author doing the meta-analysis. He might bias the conclusion from the meta-analysis by his subjective selection of studies used for the meta-analysis. If the author chooses predominantly weakly designed studies and does not weigh them down according to their weakness, the conclusions of the meta-analysis will be biased.

Conclusion

This paper has discussed a number of identification strategies that are in the current toolkit of social science researchers doing programme evaluations. It has been suggested that limited data, especially, on the counterfactual makes it essential to use the correct identification strategy so as to make causal inferences relating to the estimated treatment effects. Evaluation studies are by and large limited to studying a partial picture of reality, neglecting eventually potentially relevant interaction with other programmes and macro-level factors. Evaluation studies that explicitly refer to these partial limitations as well as to the assumptions and theoretical foundations used to motivate the choice of identification and empirical strategy are scientifically more transparent and hence enhance more interdisciplinary discussion.

The literature evaluating early childhood care and education programmes shows a number of applications of the identification strategies mentioned in this toolkit. The fact that the causal effects of many early childhood investments are not yet agreed upon by all relates on the one hand to the differences in the theoretical approach of underlying the diverse scientific disciplines doing research in this field; on the other hand, the lack of consensus is related to the differential use of identification strategies. This paper's discussion of the toolkit for *ECCE* evaluations shall enable researchers to find the most effective identification strategies in this research field. It provides an overview of the most relevant identification strategies that should be in the toolkit of each programme evaluator, in particular in the field of *ECCE* evaluations. For more elaborate discussion of the conceptual discussion and practical applications of those strategies the reader is referred to the technical and empirical sources mentioned throughout the text.

Given the renewed policy interest in the early childhood period, data collections on *ECCE* investments have been extended in many countries. In order to improve the quality of evaluation research it is important that this extension is not only composed of increasing sample sizes but also of guidance for data collections with respect to the data needs of various identification strategies. This implies, for instance, a need to account for ex-ante endowments, to document various aspects of the quality of the *ECCE*

investments and most importantly to gather sufficient information on the assignment mechanism that leads to participation in the *ECCE* treatment. A promising avenue for understanding these mechanisms would be more extensive studies of parental *ECCE* choices and their determinants, a research field which could potentially be guided by economic theories and models.

While randomized experiments are the preferred strategy for consistently identifying causal effects, they raise a number of practical and ethical questions and have some limitations regarding their external validity. However, most of those questions can be addressed by careful planning and by testing relevant control conditions. The use of data from randomized experiments can serve as an evaluation basis for other identification strategies[25]. Sometimes, randomized experiments provide unbalanced data. Assuming an exogenous assignment of treatment, propensity score matching can then be used to reach a balance of treatment and control group based on overlapping observable characteristics in those groups.

A majority of *ECCE* evaluation studies makes use of a multivariate approach to estimate treatment effects, for example, with OLS regressions. This approach provides the most direct access to estimating relationships between various *ECCE* input factors and child outcomes. However, this approach depends heavily on the availability of multiple indicators of all relevant confounding factors. As long as it cannot be rejected that there are still unobserved confounders which are not included, estimates will remain potentially spurious. Without accounting adequately for initial endowment as done first and foremost in value-added approaches, estimates are potentially biased. Family fixed effects allow to account in part for genetic predisposition as well as parental preferences, for example, with respect to *ECCE* investments and parenting styles, at least as long as those endowments can be assumed to be fixed across siblings. Twin studies provide the strongest fixes for genetic endowments while adoption studies provide unique opportunities to observe the importance of nurturing investments. However, all of these

[25] For an application of such cross-checks of identification strategies see, for instance, Heckman, Ichimura, Smith and Todd (1998) and LaLonde (1986).

fixed effects estimators demand sufficiently large samples of observations in which differences in terms of treatments are sufficiently large and all other characteristics can either be fixed or accounted for with observed information, which is challenging.

Difference-in-difference estimators provide strong identifications of causal effects and are applicable even in the case of repeated cross-sections; although, stability in unobserved characteristics between periods of observation needs to be assumed. Moreover, characteristics of the treatment and comparison groups need to be assumed being equal on average. A favourable non-experimental identification strategy is the use of instrumental variables. The instruments are potentially strong as long as external sources of information provide sufficient overlap with the occurrence of treatment and at the same time are independent from the outcome variable. Strong instruments can be used to include a 'moment' of randomization in our estimations. Policy reforms which can be seen as external events resulting in discontinuities at a certain cut-off date (or point) are prominent examples of such instrumental variables. When such an exogenous moment of randomness cannot credibly be found, the use of a quasi-structural estimation approaches might be a solution. This approach demands an explicit model that reflects the decision-making, for example, of parents to assign their child to a certain *ECCE* investment. It needs sufficient information on relevant factors determining the selection into treatment. Manski suggested a non-parametric method to get a first grasp of a treatment effect. His method uses prior knowledge about variation in the treatment to calculate bounds of the treatment effect can provide a range which the treatment effects can potentially have. However, if there is insufficient prior-information on sufficiently large sub-samples, bounds might be imprecise and cannot sufficiently be narrowed down so as to exclude the option of zero effects.

Estimated treatment effects should be interpreted in the context of the strength and weaknesses of its identification strategies. The use of meta-analysis will enhance the comparison of research results across studies and strategies. The bias from subjective choice of inclusion of studies should be limited as far as possible. Researchers who are interested in evaluating causal effects of *ECCE* investments need to pursue this quest for

causality. Combinations of different identification strategies and comparisons with studies using other strategies promise to be most effective in determining true causal effects of *ECCE* investments. The more determined researchers are in identifying the most effective strategies and the more transparent they are about the underlying assumptions, the more reliable are stated empirical results; and that in turn convinces policy makers and parents in investing into *ECCE* programmes.

Literature references

Agodini, R., & Dynarski, M. (2004). Are experiments the only option? A look at dropout prevention programs. *The Review of Economics and Statistics, 86*(1), 180-194.

Angrist, J. D., & Krueger, A. B. (2001). Instrumental Variables and the Search for Identification: From Supply and Demand to Natural Experiments. *Journal of Economic Perspectives, 15*(4), 69-85.

Armecin, G., Behrman, J. R., Duazo, P., Ghuman, S., Gultiano, S., King, E. M., et al. (2006). Early childhood development through an integrated program: evidence from the Philippines. *World Bank Policy Research Working Paper, 3922*.

Ashenfelter, O. (1978). Estimating the Effect of Training Programs on Earnings. *Review of Economics and Statistics, 60*(1), 47-57.

Baydar, N., & Brooks-Gunn, J. (1991). Effects of Maternal Employment and Child Care Arrangements in Infancy on Preschoolers' Cognitive and Behavioral Outcomes. *Developmental Psychology 27*(6), 932-945.

Becker, G. S. (1964). *Human capital; a theoretical and empirical analysis, with special reference to education*. New York,: National Bureau of Economic Research; distributed by Columbia University Press.

Becker, G. S., & Tomes, N. (1979). An Equilibrium Theory of the Distribution of Income and Intergenerational Mobility. *The Journal of Political Economy, 87*(6), 1153-1189.

Behrman, J. R., Cheng, Y., & Todd, P. E. (2004). Evaluating Preschool Programs When Length of Exposure to the Program Varies: A Nonparametric Approach. *The Review of Economics and Statistics, 86*(1), 108-132.

Bernal, R. (2007). *The Effect of Maternal Employment and Child Care on Children's Cognitive Development*. Northwestern University, U.S.A.

Bernal, R., & Keane, M. (2006). *Quasi-Structural Estimation of a Model of Child Care Choices and Child Cognitive Ability Production*. Retrieved 28.08.2008, from http://www.minneapolisfed.org/research/events/micro/bernal_keane_2006.pdf.

Bernal, R., & Keane, M. P. (2006). Child Care Choices and Children's Cognitive Achievement: The Case of Single Mothers. *IPR Working Paper, 09*.

Bertrand, M., Duflo, E., & Mullainathan, S. (2002). How much should we trust differences-in-differences estimates? *NBER Working Paper Series, 8841*.

Björklund, A., Lindahl, M., & Plug, E. (2004). Intergenerational Effects in Sweden: What Can We Learn from Adoption Data? *IZA Discussion Paper, 1194*.

Blau, F. D., & Grossberg, A. J. (1992). Maternal Labor Supply and Children's Cognitive Development. *The Review of Economics and Statistics, 74*(3), 474-481.

Blundell, R., & Macurdy, T. (1999). Labor supply: A review of alternative approaches. In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics (edition 1)* (Vol. 3, chapter 27, pp. 1559-1695): Elsevier.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Weel, B. t. (2008). The Economics and Psychology of Personality Traits. *IZA Discussion Paper, 3333*.

Borsboom, D., Mellenbergh, G. J., & Heerden, J. v. (2003). The Theoretical Status of Latent Variables. *Psychological Review, 110*(2), 203-219.

Bowles, S., Gintis, H., & Osborne Groves, M. (2005). *Unequal chances: family background and economic success*. Princeton, N.J.: Princeton University Press.

Bryson, A., Dorsett, R., & Purdon, S. (2002). *The Use Of Propensity Score Matching In The Evaluation Of Active Labour Market Policies*: Policy Studies Institute and National Centre for Social Research.

Buechner, C., & Spiess, C. K. (2007). Die Dauer vorschulischer Betreeungs- und Bildungserfahrungen - Ergebnisse auf der Basis von Paneldaten [The length of preschool care and education experiences - results based on panel data]. *DIW Discussion Papers*(687).

Campbell, F. A., Helms, R., Sparling, J. J., & Ramey, C. T. (1998). Early Childhood Programs and Success in School: The Abecedarian Study In W. S. Barnett & S. S. Boocock (Eds.), *Early Care and Education for Children in Poverty: Promises, Programs, and Long-Term Results*. Albany, NY: State University of New York Press.

Card, D. (1999). Chapter 30 The causal effect of education on earnings. In O. C. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Volume 3, Part 1 ed., pp. 1801-1863): Elsevier.

Card, D., & Krueger, A. B. (1996). Labor Market Effects of School Quality: Theory and Evidence. *NBER Working Paper, 5450*.

Cochran, W. G. (1963). *Sampling techniques* (2nd ed.). New York: John Wiley and Sons.

Cochran, W. G. (1965). The Planning of Observational Studies of Human Populations. *Journal of the Royal Statistical Society. Series A (General), 128*(2), 234-266.

Coleman, J. S. (1966). *Equality of Educational Opportunity (COLEMAN) Study (EEOS)*. Washington, DC: United States Department of Education.

Cooper, H., & Hedges, L. V. (1994). *The handbook of research synthesis*. New York: Russell Sage.

Cunha, F., & Heckman, J. J. (2007a). Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation: The University of Chicago.

Cunha, F., & Heckman, J. J. (2007b). The Technology of Skill Formation. *National Bureau of Economic Research Working Paper Series, No. 12840*.

Cunha, F., Heckman, J. J., Lochner, L., & Masterov, D. V. (2005). Interpreting the evidence on life cycle skill formation, *NBER Working Papers 11331*. Cambridge, MA: National Bureau of Economic Research.

Currie, J. (2003). When Do We Really Know What We Think We Know? Determining Causality: UCLA and NBER.

Currie, J., & Thomas, D. (1995). Does Head Start Make a Difference? *The American Economic Review, 85*(3), 341-364.

Dehejia, R. H., & Wahba, S. (2002). Propensity Score Matching Method For Nonexperimental Causal Studies. *The Review of Economics and Statistics, 84*(1), 151-161.

Duncan, G. J. (2003). Modeling the Impacts of Child Care Quality on Children's Preschool Cognitive Development. *Child Development, 74*(5), 1454 – 1475.

Dustmann, C., & Schönberg, U. (2008). The Effect of Expansions in Maternity Leave Coverage on Children's Long-Term Outcomes. *IZA Discussion Paper, 3605*.

Ermisch, J., Francesconi, M., & Pevalin, D. J. (2004). Parental partnership and joblessness in childhood and their influence on young people's outcomes. *Journal Of The Royal Statistical Society Series A, 127*(1), 69-101.

Ermisch, J. F., & Francesconi, M. (2000). The Effect of Parents' Employment on Children's Educational Attainment. *IZA Discussion Paper*(215).

Fisher, R. (1918). The causes of human variability. *Eugenics Review, 10*, 213-220.

Francesconi, M., Jenkins, S., & Siedler, T. (2005). *The Impact of Family Structure During Childhood on Later-life Attainment* (No. 1-900834-56-1): Anglo-German Foundation.

Garfinkel, I., Manski, C., & Michalopoulos, C. (1992). Micro experiments and macro effects. In I. Garfinkel & C. Manski (Eds.), *Evaluating welfare and training programs*. Cambridge, MA: Harvard University Press.

Glaeser, E. L., Laibson, D. I., Scheinkman, J. A., & Soutter, C. L. (1999). What is Social Capital? The Determinants of Trust and Trustworthiness. *NBER Working Paper, W7216*.

Gultiano, S. A., King, Elizabeth M. (2006). A Better Start in Life: Evaluation Results from an Early Childhood Development Program: Philippine Journal of Development.

Hanushek, E. A. (1974). Efficient Estimators for Regressing Regression Coefficients. *The American Statistical Association, 28*(2), 66-67.

Hanushek, E. A., & Jackson, J. E. (1977). *Statistical methods for social scientists*. San Diego, CA: Academic Press.

Harlow, H. F., & Zimmermann, R. R. (1959). Affectional responses in the infant monkey; orphaned baby monkeys develop a strong and persistent attachment to inanimate surrogate mothers. *Science, 130*(3373), 421-432.

Hausman, J. A. (1978). Specification Tests in Econometrics. *Econometrica, 46*(6), 1251-1271.

Haveman, R., & Wolfe, B. (1995). The Determinants of Children's Attainments: A Review of Methods and Findings. *Journal of Economic Literature, 33*(4), 1829-1878.

Heckman, J. J. (1976). The common structure of statistical models for truncation, sample selection and limited dependent variables and a simple estimator for such models. In *Annals of Economic and Social Measurement* (Vol. 5, pp. 475-492). New York.

Heckman, J. J. (1979). Sample Selection Bias as a Specification Error. *Econometrica, 47*(1), 153-161.

Heckman, J. J. (1990). Varieties of Selection Bias. *The American Economic Review, 80*(2), 313-318.

Heckman, J. J. (1997). Instrumental Variables: A Study of Implicit Behavioral Assumptions Used in Making Program Evaluations. *The Journal of Human Resources, 32*(3), 441-462.

Heckman, J. J. (2008a). Econometric Causality. *IZA Discussion Paper, 3425*.

Heckman, J. J. (2008b). Schools, Skills, and Synapses. *IZA Discussion Paper, 3515*.

Heckman, J. J., Ichimura, H., Smith, J., & Todd, P. (1998). Characterizing Selection Bias Using Experimental Data. *NBER Working Paper, 6699*.

Heckman, J. J., Krueger, A. B., & Friedman, B. M. (2002). *Inequality in America: what role for human capital policies?* Cambridge, Mass.: MIT Press.

Heckman, J. J., Lalonde, R. J., & Smith, J. A. (1999). Chapter 31 The economics and econometrics of active labor market programs In O. Ashenfelter & D. Card (Eds.), *Handbook of Labor Economics* (Vol. 3, Part 1, 1999, pp. 1865-2097): Elsevier.

Heckman, J. J., Lochner, L., & Taber, C. (1998). General-Equilibrium Treatment Effects: A Study of Tuition Policy. *The American Economic Review, 88*(2), 381-386.

Heckman, J. J., & Robb, R. (1985). Alternative methods for evaluating the impact of interventions : An overview. *Journal of Econometrics, 30*(1-2), 239-267.

Heckman, J. J., & Smith, J. (1999). The Pre-Program Earnings Dip and the Determinants of Participation in a Social Program: Implications for Simple Program Evaluation Strategies. *Economic Journal, 109*(457), 313-348.

Heckman, J. J., Smith, J., & Clements, N. (1997). Making the Most Out of Programme Evaluations and Social Experiments: Accounting for Heterogeneity in Programme Impacts. *The Review of Economic Studies, 64*(Special Issue: Evaluation of Training and Other Social Programmes, 4), 487-535.

Heckman, J. J., & Smith, J. A. (1995). Assessing the Case for Social Experiments. *The Journal of Economic Perspectives, 9*(2), 85-110.

Heckman, J. J., Urzua, S., & Vytlacil, E. J. (2006). Estimation of Treatment Effects under Essential Heterogeneity: University of Chicago and American Bar Foundation, University of Chicago, Columbia University.

Heckman, J. J., & Vytlacil, E. J. (2000). Local Instrumental Variables. *NBER Technical Working Paper, 252*.

Heckman, J. J., & Vytlacil, E. J. (2007). Econometric Evaluation of Social Programs, Part I: Causal Models. In J. J. Heckman & E. Leamer (Eds.), *Handbook of Econometrics* (pp. 4779-4874). Amsterdam: Elsevier.

Henry, G. T., Gordon, C. S., & Rickma, D. K. (2006). Early Education Policy Alternatives: Comparing Quality and Outcomes of Head Start and State Prekindergarten. *Educational Evaluation and Policy Analysis, 28*(1), 77-99.

Herrnstein, R. J., & Murray, C. A. (1994). *The Bell Curve The Bell Curve: Intelligence and Class Structure in American Life*. New York: Free Press.

Holland, P. W. (1986). Statistics and Causal Inference. *Journal of the American Statistical Association, 81*(396), 945-960.

Holmlund, H., Lindahl, M., & Plug, E. (2008). The Causal Effect of Parent's Schooling on Children's Schooling: A Comparison of Estimation Methods. *IZA Discussion Paper, 3630*.

Hsin, A. (2007). Mother's Time with Children and the Social Reproduction of Cognitive Skills: Department of Sociology, UCLA.

Imbens, G., & Angrist, J. D. (1994). Identification and Estimation of Local Average Treatment Effects. *Econometrica, 62*(2), 467-475.

Imbens, G., & Rubin, D. B. (2008). Causal Inferences: Harvard University.

Imbens, G., & Wooldridge, J. (2007, July 30 - August 1). Regression Discontinuity Designs. *Summer Institute 2007: What's New in Econometrics, Lecture 3* Retrieved 26.08.2008

James-Burdumy, S. N. (2005). Effect of Maternal Labor Force Participation on Child Development. *Journal of Labor Economics, 23*(1), 177-212.

Jencks, C. (1979). *Who gets ahead?: The determinants of economic success in America*. New York: Basic Books.

Karoly, L. A., Kilburn, M. R., & Cannon, J. S. (2005). *Early Childhood Interventions: Proven Results, Future Promise*. Santa Monica, CA: Rand.

Knudsen, E. I., Heckman, J. J., Cameron, J. L., & Shonkoff, J. P. (2006). Building America's Future Workforce: Economic, Neurobiological and Behavioral Perspectives on Investment in Human Skill Development. *NBER Working Paper, W12298*.

Krueger, A. B. (1999). Experimental Estimates Of Education Production Functions. *NBER Working Paper, 6051*.

LaLonde, R. J. (1986). Evaluating the Econometric Evaluations of Training Programs with Experimental Data. *The American Economic Review, 76*(4), 604-620.

Landsberger, H. (1958). *Hawthorne revisited*. New York, Ithaca: Cornell University Press.

Lefebvre, P., & Merrigan, P. (2000). The Effect of Childcare and Early Education Arrangements on Developmental Outcomes of Young Children. *Canadian Public Policy, University of Toronto Press, 28*(2), 159-185.

Leibowitz, A. (1974). Home Investments in Children. *The Journal of Political Economy, 82*(2), S111-S131.

Little, R. J. A., & Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. New York: John Wiley.

Loeb, S., Bridges, M., Fuller, B., Rumberger, R., & Bassok, D. (2005). How Much is Too Much? The Influence of Preschool Centers on Children's Social and Cognitive Development. *National Bureau of Economic Research Working Paper Series, No. 11812*.

Ludwig, J., & Miller, D. L. (2006). Does Head Start Improve Children's Life Chances? Evidence from a Regression Discontinuity Design. *IZA Discussion Paper, 2111*.

Magnuson, K., & Waldfogel, J. (2005). Child Care, Early Education, and Racial/Ethnic Test Score Gaps at the Beginning of School. *The Future of Children, 15*(1), 169-196.

Magnuson, K. A., Meyers, M. K., Ruhm, C. J., & Waldfogel, J. (2004). Inequality in Preschool Education and School Readiness. *American Educational Research Journal, 41*(1), 115-157.

Manski, C. F. (1989). Anatomy of the Selection Problem. *The Journal of Human Resources, 24*(3), 343-360.

Manski, C. F. (1990). Nonparametric Bounds on Treatment Effects. *The American Economic Review, 80*(2), 319-323.

Manski, C. F. (1993). Identification Problems in the Social Sciences. *Sociological Methodology, 23*, 1-56.

Mayer, S. E. (1997). *What money can't buy: family income and children's life chances*. Cambridge, Mass.: Harvard University Press.

Moffitt, R. A. (2004). The Role of Randomized Field Trials in Social Science Research - A Perspective from Evaluations of Reforms of Social Welfare Programs. *American Behavioral Scientist, 47*(5), 506-540.

Olds, D., Henderson, C. R. J., Cole, R., Eckenrode, J., Kitzman, H., Luckey, D., et al. (1998). Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior: 15-Year Follow-up of a Randomized Controlled Trial. *Journal of the American Medical Association, 280*(14), 1209-1288.

Pavlov, I. P. (1927). *Conditioned Reflexes: An Investigation of the Physiological Activity of the Cerebral Cortex (translated by G. V. Anrep)*. London: Oxford University Press.

Plug, E., & Vijverberg, W. (2001a). Schooling, Family Background, and Adoption: Does Family Income Matter? *IZA Discussion Paper, 246*.

Plug, E., & Vijverberg, W. (2001b). Schooling, Family Background, and Adoption: Is it Nature or is it Nurture? *IZA Discussion Paper, 247*.

Puhani, P. (2000). The Heckman Correction for Sample Selection and Its Critique - A Short Survey. *Journal of Economic Surveys, 14*(1), 53-68.

Ramey, S. L., & Ramey, C. T. (2003). Understanding Efficacy of Early Educational Programs: Critical Design, Practice, and Policy Issues (chapter 2). In A. J. Reynolds, M. C. Wang & H. J. Walberg (Eds.), *Early Childhood Programs for a New Century* (pp. 35-70). Washington DC: Child Welfare League of America.

Ravallion, M. (2005). Evaluating Anti-Poverty Programs. *World Bank Policy Research Working Paper, 3625*.

Rosenbaum, P., & Rubin, D. B. (1984). Reducing Bias in Observational Studies Using Subclassification on the Propensity Score. *Journal of the American Statistical Association, 79*, 516-524.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika, 70*(1), 41-55.

Rosenzweig, M. R., & Wolpin, K. I. (1995). Sisters, Siblings, and Mothers: The Effect of Teen-Age Childbearing on Birth Outcomes in a Dynamic Family Context. *Econometrica, 63*(2), 303-326.

Rubin, D. B. (1974). Estimation of causal effects in randomized and non-randomized studies. *Journal of Educational Psychology, 66*, 688-701.

Ruhm, C. J. (2000). Parental Employment and Child Cognitive Development. *NBER Working Papers, 7666*.

Rutter, M. (2002). The Interplay of Nature, Nurture, and Developmental Influences - The Challenge Ahead for Mental Health. *Arch Gen Psychiatry, 59*, 996-1000.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2002). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 8a - Measuring the Impact of Pre-School on Children's Cognitive Progress over the Pre-School Period*. London: Institute of Education.

Sammons, P., Sylva, K., Melhuish, E., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2003). *The Effective Provision of Pre-School Education (EPPE) Project: Technical Paper 8b - Measuring the Impact of Pre-School on Children's Social/behavioural Development over the Pre-School Period*. London: Institute of Education.

Schultz, T. W. (1961). Investment in human capital. *American Economic Review, 51*(1), 1-17.

Schweinhart, L. J., Barnes, H. V., & Weikart, D. P. (1993). *Significant Benefits: The High/Scope Perry Preschool Study through Age 27 (Monograph 10)*. Ypsilanti, MI: High/Scope Educational Research Foundation.

Shonkoff, J. P., & Phillips, D. A. (2000). *From Neurons to Neighborhoods: The Science of Early Childhood Development*: National Acedemy Press.

Stock, J. H., Wright, J. H., & Yogo, M. (2002). A Survey of Weak Instruments and Weak Identification in Generalized Method of Moments. *American Statistical Association, Journal of Business & Economic Statistics, 20*(4), 518-529.

Sylva, K., Melhuish, E., Sammons, P., Siraj-Blatchford, I., Taggart, B., & Elliot, K. (2003). *The Effective Provision of Pre-School Education (EPPE) Project: findings from the pre-school period, brief number RBX15-03*. London: DfES Publications.

Theil, H. (1953). Estimation and simultaneous correlation in complete equation systems. Mimeo. The Hague: Central Planning Bureau.

Todd, P., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function Cognitive Achievement. *The Economic Journal, 113*(F3-F33).

Todd, P., & Wolpin, K. I. (2004). The Production of Cognitive Achievement in Children: Home, School and Racial Test Score Gaps. *PIER Working Paper, 04*(019).

Vandell, D. L., & Ramanan, J. (1992). Effects of early and recent maternal employment on children from low-income families. *Child Development, 63*, 938-949.

Waldfogel, J., Han, W.-J., & Brooks-Gunn, J. (2002). The Effects of Early Maternal Employment on Child Cognitive Development *Demography, 39*(2), 369-392.

Wong, V. C., Cook, T. D., Barnett, W. S., & Jung, K. (2008). An Effectiveness-Based Evaluation of Five State Pre-kindergarten Programs. *Journal of Policy Analysis and Management, 27*(1), 122-154.

Würtz, A. (2007a). Allocation of Parental Time and the Long-Term Effect on Children: Department of Economics and CIM, Aarhus School of Business, University of Aarhus.

Würtz, A. (2007b). The Long-Term Effect on Children of Increasing the Length of Parents' Birth-Related Leave. *Dept. of Economics and CIM, Aarhus School of Business, Working Paper, 07-11*.

Maastricht Graduate School of Governance
Working Paper Series

List of publications

2008

No.	*Author(s)*	*Title*
001	*Roelen, K. and Gassmann, F.*	Measuring Child Poverty and Well-Being: a literature review
002	*Hagen-Zanker, J.*	Why do people migrate? A review of the theoretical literature
003	*Arndt, C. and C. Omar*	The Politics of Governance Ratings
004	*Roelen, K., Gassmann, F. and C. de Neubourg*	A global measurement approach versus a country-specific measurement approach. Do they draw the same picture of child poverty? The case of Vietnam
005	*Hagen-Zanker, J. M. Siegel and C. de Neubourg*	Strings Attached: The impediments to Migration
006	*Bauchmüller, R.*	Evaluating causal effects of Early Childhood Care and Education Investments: A discussion of the researcher's toolkit

2007

No.	*Author(s)*	*Title*
001	Notten, G. and C. de Neubourg	Relative or absolute poverty in the US and EU? The battle of the rates
002	Hodges, A. A. Dufay, K. Dashdorj, K.Y. Jong, T. Mungun and U. Budragchaa	Child benefits and poverty reduction: Evidence from Mongolia's Child Money Programme
003	Hagen-Zanker, J. and Siegel, M.	The determinants of remittances: A review of the literature
004	Notten, G.	Managing risks: What Russian households do to smooth consumption

005	Notten, G. and C. de Neubourg	Poverty in Europe and the USA: Exchanging official measurement methods
006	Notten, G and C. de Neubourg	The policy relevance of absolute and relative poverty headcounts: Whats in a number?
007	*Hagen-Zanker, J. and M. Siegel*	A critical discussion of the motivation to remit in Albania and Moldova
008	*Wu, Treena*	Types of Households most vulnerable to physical and economic threats: Case studies in Aceh after the Tsunami
009	*Siegel, M.*	Immigrant Integration and Remittance Channel Choice
010	*Muñiz Castillo, M.*	Autonomy and aid projects: Why do we care?

2006

No.	*Author(s)*	*Title*
001	Gassmann, F. and G. Notten	Size matters: Poverty reduction effects of means-tested and universal child benefits in Russia
002	Hagen-Zanker, J. and M.R. Muñiz Castillo	Exploring multi-dimensional wellbeing and remittances in El Salvador
003	Augsburg, B.	Econometric evaluation of the SEWA Bank in India: Applying matching techniques based on the propensity score
004	Notten, G. and D. de Crombrugghe	Poverty and consumption smoothing in Russia

2005

No.	*Author(s)*	*Title*
001	Gassmann, F.	An Evaluation of the Welfare Impacts of Electricity Tariff Reforms And Alternative Compensating Mechanisms In Tajikistan
002	Gassmann, F.	How to Improve Access to Social Protection for the Poor? Lessons from the Social Assistance Reform in Latvia