Interpreting program evaluation studies of the impact of ECEC on child development

Jeff Borland
Department of Economics
University of Melbourne
May 2023

Submission to the Productivity Commission Inquiry into 'Early Childhood Education and Care'

Introduction

These notes provide commentary on empirical research on the impact of Early Childhood Education and Care (ECEC) on child development. The commentary is in the form of general principles that I argue should inform a reading of the available research. A more traditional submission would address a question such as: 'What is the evidence on the impact of ECEC on child development?'. However, over the past several years, in making presentations of research related to ECEC, through being asked to review literature on the impacts of ECEC, and in listening to commentary on empirical research on ECEC at various conferences, I've come to see a need for greater emphasis on using appropriate approaches for interpreting the findings from that research. This submission therefore seeks to provide a brief overview of those approaches.

The touchstones

The touchstones for interpreting and applying findings from program evaluation studies are internal validity and external validity. **Internal validity** concerns whether a study identifies a credible causal impact of a program or policy. **External validity** is about the reasonable scope of application of a study's findings on the impact of a program or policy; for example, the types of populations or the types of situations for which the findings of the study are thought to be relevant.

Internal and external validity can be thought of as a two-stage check on the value of a program evaluation study: Policy design should only rely on findings from program evaluation that are internally valid. But findings on program impacts meeting that standard, must also be relevant to the policy design issue under consideration; that is, externally valid.

<u>Internal validity</u>

In creating an overall evidence base, it's critical to weight program evaluation studies according to the extent to which they identify causal impacts of ECEC programs. Estimated causal impacts tell us about how a program affects an outcome of interest in a way that would not have happened in the absence of the program. It seems automatic that this information is what we want to guide policy making; a knowledge of the 'pure' effect of a program or policy.

An estimated program impact will not be causal where it confounds the 'pure' or causal impact of a program with other factors that also determine the outcome of interest. For example, suppose participation in an ECEC program for children from disadvantaged backgrounds is voluntary. Then any estimated impact of that program is likely to reflect both the causal impact of the program, but also other factors that are correlated with the decision by a child's primary caregiver to agree to participation (which might include parental education, access to transport and family income). Because an estimated impact that is non-causal confounds the causal impact of the program with these other factors, it is much less valuable for telling us about the impact on child development of the program being studied.

Judging whether a study has identified a credible causal impact of a program is primarily about the extent to which influences on child development, apart from the program itself, have been controlled for. Whether an estimated program impact can be considered as credibly causal therefore depends on a study being able to make a convincing case that other factors that might have caused the outcome of interest to differ between the program participants and control group, have been successfully controlled for.

This is not a 'zero/one' exercise. In any program evaluation study, there is likely to be some residual doubt about the comparability of program participants and the control group. Instead, the question is about the degree to which the study is judged likely, in estimating the program's impact, to have controlled for other sources of differences in outcomes between the program participants and control group.

Making that judgement should incorporate a range of considerations. The methodology used is one important consideration, but should not be the sole consideration. For example, randomised controlled trials (RCTs) have natural advantages for identifying causal effects over methods such as 'before/after'; however, it's also the case that much can go wrong with a RCT, so that it's important to judge any study on how well it applies its chosen methodology (see, for example, the early debate between Burtless, 1996 and Heckman and Smith, 1996).

Ultimately, what is required in assessing the internal validity of any study is to think through the set of potential 'other factors' that might cause different outcomes between program participants and the control group; and then assess to what extent those factors have been successfully controlled for. The factors could relate to differences in characteristics of children who have participated in an ECEC program and the control group; or could relate to other issues that are more difficult to observe.

An example of the former factor would be where a control group of children differs from program participants in characteristics such as average age or family background. Differences in characteristics are particularly relevant as a potential confounder of the estimated impact of a program where: (i) The basis of selection into the sample of program participants and the control group departs substantially from random selection; and (ii) In estimating the program impact, no attempt is made to control for how differences in characteristics of program participants and the control group might have affected the outcome being studied (see, for example, Newman et al., 2022).

An example of the latter factor might be data quality and measurement. Suppose that a study has involved collection of data on children participating in a program and a control group, by researchers who know from which group they are collecting data. Then it is conceivable that the data collected on outcomes for program participants relative to the control group will be affected by that knowledge.[1]

External validity

The applicability of the findings from any program evaluation study depends critically on the specific details of the study. This is even more the case for evaluation studies of ECEC programs, notable for the high degree of variability in estimated impacts on child development. Key mediating factors in the variability of estimated impacts are known to be: (i) the level of disadvantage of a child's family (income, risk factors etc); (ii) age of child; (iii) type of program (for example, targeted versus universal); and (at least according to some commentaries) (iv) the quality of program being studied. Hence, it's important to be clear about what aspect of knowledge on the impact of ECEC programs that a study is contributing to.

---

[1] As a further example on this point, Heckman et al. (1998) investigated what were the critical aspects in controlling for confounding factors in studies of the impact of active labour market programs on employment outcome. They highlighted three main factors: first, having data on both program participants and the control group from the same data source; second, for having program participants and control group members from the same geographic region; and third, being able to control for a relatively rich set of background characteristics relevant to determining employment status.

A common example of not abiding by this principle is interpreting estimated impacts from ECEC programs designed for children from highly disadvantaged backgrounds as telling us about the likely impact of universal ECEC programs, as pointed out by Michael Baker (2011, p.1069) in his 2011 Innis Lecture:
'Policy makers and advocates often cite the research on targeted programs in support of universal programs, although this is problematic for a number of reasons.'

A more subtle example concerns thinking about the impact of ECEC programs by the level of disadvantage of children participating in those programs. A recent survey (Duncan et al., 2022, p.1) has characterised evidence on the impact of ECEC programs targeted at children from disadvantaged backgrounds as 'mixed'. This appears to contrast, for example, with findings from the study of the Australian Early Years Education Program (EYEP) that I have been involved with, which finds large impacts for children who are living in extreme adversity (Tseng et al., 2022). One explanation for the apparent difference in conclusions is that different definitions of disadvantage are being used. The programs that Duncan et al. (2022) define as targeted, have as participants children whose family backgrounds extend over a broad range of levels of disadvantage; whereas the EYEP study was highly targeted at a group of children at the upper end of the spectrum of disadvantage. Hence, it may be that the findings are reflecting that targeted programs can have large impacts for children from extremely disadvantaged backgrounds, but more limited impacts across the full spectrum of what would be regarded as disadvantage. This example illustrates the importance of understanding in detail the characteristics of children participating in a program (and indeed the program's design), in order to interpret findings on the impact of a program, and to properly appreciate the implications for policy design.

At the same time, it's also important that deciding on the external validity of a program evaluation study is not regarded as 'zero/one' (similar the point I made earlier regarding internal validity). Finding studies that are a 'perfect match' for a new policy problem is likely to be difficult, if not impossible. For example, consider the policy problem of extending subsidised ECEC within a specific geographic region to a broader population of children than are currently covered by that program. Just by the fact that the policy involves covering a new population within a specific region, it is unlikely that a study exists which tells us the impact of extending the same type of ECEC program proposed, to exactly the type of population who will be affected.

The task of using evidence to guide policy design in ECEC is therefore to make judgements about the appropriate weight to attach to existing program evaluation studies, depending on the degree of 'match' between the policy design issue under consideration and the features of those studies (such as type of program, type of population and general policy environment).

In making judgements about the relevance of existing research studies in this way, I would argue that an important consideration is the volume of evidence. As an example, consider the relevance of the evidence on the impact of universal care on child development outcomes from international research. One study, done for a different country, that comes to a particular conclusion, could be appropriately viewed as not highly relevant for Australia, due to differences such as the type of program or institutional features. However, multiple studies done for a variety of different countries, which come to similar conclusions, and where those studies cover a variety of types of programs and populations, could justifiably be argued to be identifying systematic findings, that do then become relevant. Such could be argued to be the case with respect to conclusions from international studies of universal ECEC regarding the varying impact on child development depending on the level of disadvantage in family background (see, for example, van Huizen and Plantenga, 2018).

Interpretation of the policy relevance of the estimated impact of a program also needs to be based on the type of impact identified. Table 1 below lists some of the types of impacts that can be identified in program evaluation studies. For purposes of policy design, it's essential to understand that the alternative impacts are: (i) Of differing relevance depending on the policy design issues being addressed; and (ii) Likely to differ in size due to heterogeneity in the impacts of any ECEC program between different potential groups of children who might participate.

Addressing these points in reverse order: The different types of impact – such as shown in the table below - are estimated for different subsets of the population of potential program participants. Whenever the impact of a program varies between individuals in the population, it follows that the different types of impacts can vary in size (Heckman, 2001). For example, suppose the benefits of an ECEC program vary inversely with family income, and that the program is currently provided to children living in families with income less than $50,000. Then the estimated average impact estimated for all children currently participating (average effect of treatment on the treated) will be greater than the impact estimated, for example, for children in families with income between $50,000 and $55,000 (marginal average effect of treatment). Furthermore, any specific estimated impact may or may not be relevant, depending on the policy design issue of interest. Continuing with the example just used, if the policy question is whether to continue with the program targeted at children living in families with incomes less than $50,000, then it is the average effect of treatment on the treated that is policy-relevant; whereas if the policy question is whether to expand the program to children living in families with income between $50,000 and $55,000, then it is marginal average treatment effect that is policy-relevant.

As another example, in the study of the Early Years Education Program with which I have been involved, for reasons of data availability, the impact we estimate is for children who participate in the program for more than a threshold number of days (the average effect of

treatment on the treated). Arguably, what is more relevant for policy design, is knowing the impact on the whole population of children who it was intended would participate (intention to treat); as this effect incorporates into the estimated impact that, was the program to be rolled out on a larger scale, it may be difficult to engage some children and their families with the program.

Table 1: Types of impact estimates from program evaluation studies

Impact	Population estimated on	Relevant for policy questions about
Average effect of treatment	Whole potential population of participants	Should the program be made available to the whole population?
Average effect of treatment on the treated	Population who actually participate in a program	Is the program benefiting those who participate?
Intention to treat	Population who were intended to participate in a program	Is the program benefiting the population who are intended to participate?
Marginal average effect of treatment	Population who would stop/begin participating in a program with a marginal change in eligibility conditions	Should the program be contracted/expanded?

Also critical in interpreting findings from studies of ECEC programs, is to be aware that estimated impacts from program evaluation studies tell us about outcomes for the treatment group **relative** to a control group.

The estimated impact of a program is the difference between outcomes for the treatment group and a control group. Hence, when making comparisons of estimated impacts from studies across different time periods or between different groups of children, those estimates can vary, either because the effect of program on the treatment group is different, or because the outcome for the control group is different. Put another way, a difference in estimated impacts does not necessarily imply that outcomes for participants in the program changed or were different; it may instead be variation in the outcome for the control group that is causing the different estimated impact.

An important implication is that the same program, applied to different populations of children or in different geographic areas, can have different impacts, due to variation in access to and in the quality of ECEC for children who do not participate in the program. One example is how a decreasing size of impact over time of ECEC programs targeted at disadvantaged children in the United States has been interpreted to, at least partly, reflect an improving quality of ECEC for those children (Duncan and Magnuson, 2013, p.114).

Another example is the suggestion that impacts of targeted programs may vary between less and more disadvantaged children, not just because the latter group might benefit more from compensatory effects on development of a targeted program, but also because the quality of alternative care they would receive in the absence of participating in the program (either centre or home-based) would be lower (Duncan et al., 2022, pp.40-41).

As a final point, it's worth noting that, while evidence is growing, there is still a huge amount that we don't know about the impact of ECEC on child development. As a few examples:
• (i) What are relative impacts of differences in quantity and quality of education and care on the estimated impact of ECEC on child development?;
• (ii) What explains fade-out that has been found to occur in the impact of some ECEC programs in the years after children complete the programs?;
• (iii) How should differences in the quality of ECEC and the mechanics of how quality affect child development be thought about? (for example, Duncan et al., 2022);
• (iv) What are the channels by which ECEC programs have an impact on child development (for example, via impacts and inter-linkages between cognitive and non-cognitive development)?

References

Baker, M. (2011), 'Innis Lecture: Universal early childhood interventions: What is the evidence base?', *Canadian Journal of Economics*, 44(4): 1069-1105.

Burtless, G. (1995), 'The case for randomized field trials in economic and policy research', *Journal of Economic Perspectives*, 9(2), 63-84.

Duncan, G., A. Kalil, M. Mogstad and M. Rege (2022), 'Investing in early childhood development in preschool and at home', National Bureau of Economic Research Working Paper no. 29985.

Duncan, G.J. and K. Magnuson (2013), 'Investing in preschool programs', *Journal of Economic Perspectives*, 27(2), 109-32.

Heckman, J. (2001), 'Micro data, heterogeneity and the evaluation of public policy: Nobel lecture', *Journal of Political Economy*, 109, 673-748.

Heckman, J., H. Ichimura, J. Smith and P. Todd (1998), 'Charaterizing selection bias using experimental data', *Econometrica*, 66(5), 1017-98.

Heckman, J. and J. Smith (1995), 'Assessing the case for doing social experiments', *Journal of Economic Perspectives*, 9(2), 85-110.

Newman, S, J. McLoughlin, H. Skouteris, C. Blewitt, E. Melhuish and C. Bailey (2022), 'Does an integrated wrap-around school and community service model in an early learning setting improve academic outcomes for children from low socioeconomic backgrounds?', *Early Child Development and Care*, 192: 816-30.

Tseng, Y., Jordan, B., Borland, J., Clark, M., Coombs, N., Cotter, K., Guillou, M., Hill, A., A. Kennedy and J. Sheehan (2022), *Changing the Life Trajectories of Australia's Most Vulnerable Children – Report no.5: 36 months in the Early Years Education Program: An assessment of the impact on children and their primary caregivers*; accessed at: https://fbe.unimelb.edu.au/__data/assets/pdf_file/0006/4360173/9597-MU-EYEP-Report-5-WEB52.pdf

Van Huizen, T. and J. Plantenga (2018), 'Do children benefit from universal early childhood education and care? A meta-analysis of evidence from natural experiments', *Economics of Education Review*, 66: 206-22.