“The evaluation of the School Support Program under DEIS”: An assessment

Kevin Denny

UCD School of Economics and Geary Institute for Public Policy

Geary PPR2015.01
April 2015
Executive summary

The note discusses a recent evaluation by Weir & Denner (2014) of the School Support Program, a key component of the DEIS (Delivering Equality of Opportunity in Schools) initiative. DEIS is the government’s main initiative to tackling socio-economic disadvantage. The evaluation considered here, part of a wider set of evaluations by the Education Research Center focuses on urban primary schools only. I discuss the evaluation with regard to two criteria (i) identification: whether the study design in principle allows one to assess the efficacy of the program and (ii) estimation: whether the statistical analysis, as presented, is well carried out. I argue that, by both criteria, there are serious deficiencies in the analysis. This means that one cannot conclude that the programme is a success or otherwise. A fundamental problem is that the design of the evaluation did not include any control group contrary to best evaluation practice.
Introduction
Education in Ireland is characterized by high and persistent levels of inequality at all levels. That is, outcomes vary significantly with socio-economic status. At primary and secondary level this means that academic achievement depends to a significant degree on one’s background. Progression to third level education is heavily skewed towards those from better off backgrounds. These broad patterns are well known although the mechanisms through which these inequalities exist - and persist - are not nearly as well understood. It is very likely that there are complex reasons for the under-achievement of children from less well-off backgrounds including financial constraints, attitudes to education and school quality.

It is welcome, therefore, that the government has introduced a number of interventions to address these inequalities. The Delivering Equality of Opportunity in Schools (DEIS) programme launched by the Department of Education and Skills (then the Department of Education and Science, henceforth DES) in May 2005 is the primary policy intervention to address the educational needs of children and adolescents from disadvantaged communities at primary and secondary level.

Here, I consider an evaluation of the programme for urban primary schools which the Education Research Center (ERC) was commissioned to carry out by DES, beginning in 2007. The report was prepared by Susan Weir and Sylvia Denner. This present document does not attempt to exhaustively describe the evaluation but focuses on what I think are the key features.

The term “evaluation” can mean different things so in assessing an evaluation it is important to be clear about the criteria one is applying. My understanding is that this report is designed to measure the effect of the program on academic achievement by children in those schools which participate in the School Support Programme (SSP). To know exactly what the effect is one would need to know the counterfactual: what would have happened if the same children had not participated in the program. This is not possible since nobody can both experience and not experience the program. So, as noted by Shadish, Cook and Campbell (2002, p5) “..a central task for all cause-probing research is to create a reasonable approximation to this physically impossible counterfactual”. There are several different ways of doing this each with their own strengths and weaknesses and the choice of technique depends on scientific, practical and also ethical considerations.
Nonetheless, without any counterfactual one simply has no way of estimating the effect of the intervention. For most purposes, a counterfactual is based on a control group: individuals who are comparable to those who experience the intervention but who did not.

To get some sense of the importance of the control group, Figure 1 shows two potential scenarios which is comparable to some of the figures in the evaluation. The green line shows what happens to the outcome of the participants over time. The time dimension should refer to what happened before as well as after the program was introduced. That the line is upward sloping might be interpreted as implying that the program is a success. In the absence of “before” data however it may well be the case that this group were improving already. Consider two possible controls groups. The red line shows what is happening to group 1. Since they are also improving, for some reason, the effect of the program (the difference between the green and red lines) may be quite small. If however the control group is getting worse over time (group 2, the blue line) then the effect of the programme is much bigger. Either scenario is possible although the first (group 1) is probably more likely and the absence of a control group is very likely to exaggerate the effect of the programme.

![Figure 1: The role of a control group](image)

In some cases, an evaluation is commissioned after an intervention has been initiated or even completed. In such a case, the evaluators are required to construct a counterfactual ex-post which
may be difficult or impossible if the appropriate data hasn’t been collected from the start. Denny et al. (2010, 2014) is an evaluation of UCD’s access program for students from socio-economically disadvantaged backgrounds. In that case they were able to construct a control group ex-post by taking advantage of the roll-out of the programme over time. It is not ideal. Best practise is to design an evaluation at the same time as the intervention so one can choose a control group and collect the necessary data. Johnson et al. (2014) is a good example of a high quality evaluation looking at one school in the United States which used matching techniques to construct a comparison group. In the next section I give a brief overview of the evaluation while section 3 is my commentary on the data analysis.

Description
The research involved collecting data in three years 2007, 2010 and 2013. The intervention started in academic year 2006/07. The first wave of data (collected in early 2007) represented “baseline data”. At that point, pupils would have already experienced the programme (presumably for about six months) so it cannot be regarded as measure of pre-programme achievement – nor do the authors claim it to be.

The structure of the data is somewhat complicated but is clearly explained. In 2007, pupils in 2nd & 3rd classes and 6th class were tested. In 2010 the 2nd and 3rd class from the 2007 cohort (then in 5th & 6th class respectively) were tested as well as new pupils in 2nd and 3rd class. In 2013 the latter two groups (now in 5th & 6th class respectively) as well as new pupils in 2nd and 3rd class were tested. The cohort from 2007 would have completed primary school at that stage.

120 schools were sampled and each school participated in all three waves. It is not clear how these were selected, for example were they randomly chosen. It would be helpful if the report indicated how many DEIS primary schools as well as how many primary schools of all types there are. From the DEIS website, it appears there are about 3200 mainstream primary schools of which 658 are in DEIS. Of these 336 are urban so the sample studied is about 36% of the relevant population of schools. For each class (2nd, 3rd etc) in each year about 4300 pupils were tested. It would have been helpful to know how many individual classes were tested and what the range was between schools e.g. did some schools provide multiple 2nd classes for testing or was one the norm?
The report focuses on the School Support Programme (SSP) which entitles participants to a range of supports and access to other programmes. There does not appear to be any information in the report on what exactly is involved in these programs or the take-up. Even if this information is published elsewhere it would be helpful for the reader to have some idea of what the program involves or at least provide a reference to an appropriate source.

DEIS schools are in two bands, with schools in band 1 receiving more support than in band 2, specifically to support lower pupil-to-teacher ratios. Schools are assigned to bands on the basis of having greater concentrations of disadvantage. It would very useful if the report were to set out the criteria for a school to qualify as a DEIS school and in turn how exactly schools are divided between bands 1 and 2. It would also be helpful to know also whether schools could decline the offer of participation and, if so, whether any did. It is possible that a school might not wish to be perceived as “disadvantaged”. As will be discussed below, the issue of selection into the treatment is critical to drawing conclusions about its efficacy.

The core of the data analysis here consists of tables and graphs. The tables show firstly the average raw scores (in reading and mathematics, separately) for each class (e.g. all 2nd class pupils) for the three years. Secondly, they show the percentage performing below the 10th and above the 90th percentiles, based on test scores of a national sample on whom the tests were standardized. Using distributionally sensitive measures such as these is a good idea since focusing on the mean can be misleading: one may want to pay particular attention to those at the low end of the distribution. However, no information on these national tests is provided such as how the sample was selected, when these tests were administered, the sample size and so on. If this group includes DEIS schools then their performance and that of DEIS schools will be correlated by construction, so it is unclear what the comparison is telling us. For reading attainment these results are in tables 5-7 of the report. Table 8 breaks down the results by bands 1 and 2. As one would expect, pupils in band 1 (the more disadvantaged) do consistently worse than band 2. Both increase over time at about the same rate although there is no test for any differences.

The pattern in the results is generally very clear: the absolute performance of a given class (e.g. all 3rd class pupils) increases over time. Relative to the national norms, the performance also increases over time for each class. For example, from Figure 2 one can see that with a national norm of 100
(in each year), 2\textsuperscript{nd} class pupils in DEIS schools were scoring on average about 92.5 in 2007, 95 in 2010 and 97 in 2013.

While most of the analysis is cross-sectional there is a brief discussion of longitudinal change as the performance of the same group of pupils is tracked over time. The authors point out that there is a small increase in the scores (relative to the national norm, so allowing for the fact that pupils’ attainment generally improve as they get older) and that this change is “statistically significant”.

The description of the mathematics testing parallels that of the reading tests and the results are very similar.

\textbf{Commentary}

In assessing whether a comparison such as this evaluation provides a rigorous assessment of the impact of the programme there are two distinct criteria to be used: \textit{identification} and \textit{estimation}. Identification refers to whether one has the right type of data to allow one to identify, in principle, the effect of the programme in question. Estimation refers to the statistical procedures used (and their presentation) to draw conclusions from the sample.

\textit{a. Identification}

As discussed in the introduction, in any programme evaluation it is essential to have a reliable estimate of the counterfactual: what would have happened if the participants had not been part of the programme. As the authors note in their conclusions the absence of such a comparison group is a major limitation. Given the importance of this, it would be more appropriate to explain this at the outset of the study. The authors’ defence of the absence of a control group is that “\textit{However, the outcomes of a series of national assessments in reading and mathematics that have taken place over the past couple of decades indicate that no major changes in reading standards have occurred}” (p 21).

No other information is provided on these assessments. So we do not know, for example, when they occurred, whether the same instruments were used, what schools were tested and what “major change” constitutes since some of the differences observed in the study under consideration are, arguably, small. The absence of any source of reference to these assessments is a serious deficiency. Essentially we are asked to take on trust, in a sense which is undefined, that
there were no changes in attainment levels generally. What has happened over the *past couple of decades* is irrelevant: one needs to know about what has happened over the *last few years* i.e. since the introduction of the programme.

Let us assume one knows what happens to attainment nationally using comparable tests over the same period (2007-2013). One might be tempted to compare the rate of change in the DEIS schools with the change nationally. There are several reasons why this is invalid. Firstly, and most obviously, the national results will include the DEIS schools so this comparison will tend to bias downwards the apparent effect of the programme on DEIS schools. Secondly, even if one had managed to estimate the change in non-DEIS schools, a comparison between the change in these two groups is likely to be unhelpful. Simply comparing the averages of participants and non-participants (known in statistics as the *average treatment effect*) is potentially misleading. In the context of evaluating a job-training program, James Heckman (1997), the Nobel prize winning economist said this is like “picking a millionaire at random to participate in a training program for low-skilled workers”. In other words, *the comparison group should be comparable*: fee-paying schools catering for children from well-off backgrounds are not.

This gets to the heart of the matter: to evaluate the effect of the program one has to compare participants with pupils who are as similar as possible but who did not participate. Pupils who are from affluent communities, with well-off parents and most likely in better resourced schools will differ in ways that we do not observe and therefore should not be in the comparison group. The same argument applies at the other extreme, children who are the most disadvantaged are not representative of the programme participants. To see how the problem arises and how it might be solved consider Figure 2 below.

Along the horizontal axis is the measure of socio-economic advantage which is presumed to determine entry to the programme. Schools in areas which are sufficiently *disadvantaged* (below $X_0$) participate and not otherwise. Some measure of attainment (which could be a level or a rate of change) is on the vertical axis. The upward sloping lines reflect the fact the more disadvantaged schools do worse on average but the DEIS program, by assumption, raises the achievement of those schools which participate.
Comparing the average DEIS school with the average non-DEIS school is misleading. Instead, one should compare those schools which just qualified for DEIS with those who just failed to qualify – as they were insufficiently disadvantaged. This difference, the red vertical arrow, is a valid measure of the effect of the programme. The green dashed line shows how the DEIS schools would have performed if the program did not exist (or before it existed). Comparing the averages of the two groups is likely to a downward biased estimate of the effect.

This approach is known as a regression discontinuity design (RDD) and has been used extensively in education and economics research\(^1\). Even where the criterion for programme participation is not as sharp as indicated in the diagram (so called fuzzy-design) the method can still be used. It seems to me that this research design (which can be thought of a quasi-experimental) is simple to implement, feasible and would have been very informative in this context. RDD is not the only method available and there is a vast literature on statistical evaluation techniques.

---

\(^1\) The first application was by education researchers Thistlewaite & Campbell (1960). It has since become widely used in economics, epidemiology and other disciplines. See for example Lee & Lemieux (2010)
Obviously, it is not possible to use RDD (or any other standard technique) when there is no comparison group available. However I think, in principle, the child cohort of the *Growing Up in Ireland* (GUI) study could be used if the criterion for assigning schools to the DEIS program was embedded in the data - that is whatever the Department of Education used for the horizontal axis in the above diagram. Since the GUI is a superb source of information on school characteristics as well as characteristics of the children, their parents and teachers, this would allow a much more powerful analysis of the determinants of educational attainment which would greatly enhance the analysis of the DEIS program. Such a study would be inexpensive and make good use of an important investment in research infrastructure. Indeed, academic researchers are likely to do this at no cost to the Exchequer if the data is available.

One final issue here is that the evidence consists of both cross-sectional comparisons (for example comparing 2nd class students in one year with 2nd class students three years later) and longitudinal ones (following the same group of pupils over time). Presumably, these are answers to different questions but it is not specified what those questions are and which, if either, is telling us whether the program works.

**b. Estimation**

The results shown in the report are potentially interesting but without any details it is impossible to assess. For example, to determine whether a difference is significantly different (i.e. is not just due to sampling variation) depends on a range of factors. One needs to know at what level the test is conducted - a critical p-value of .05 is common but not universal – and, indeed, what test was used, and what the actual value of the test was. With large samples (and hence more powerful tests) there is an argument for setting the bar higher - that is setting the critical value *lower*. In this case, it is not straightforward because one needs to take into account the hierarchical (as well as the longitudinal) nature of the data: pupils are in classes and in schools so one needs to allow for potential clustering. For example a sample of 3000 pupils in 120 classes in 100 schools can be expected to have much less independent variation than the same number that are completely randomly selected. Given that the comparisons are relative to a national norm which in itself is based on a (presumably random) sample, it is unclear how the comparisons were made. That is, was the sampling variation in the norm group taken into account?
It may be the case, that this evaluation was not deemed the appropriate place for these statistical details and that these are provided elsewhere (this is not indicated). In my opinion, this is not good practise. It is possible (and indeed common) in evaluations to provide sufficient statistical details even if the intended audience is not a technical one through the use of footnotes and technical appendices. A relevant comparison is the evaluation of University College Dublin’s access program by Denny et al. (2009) which, although written for the non-academic reader, includes extensive statistical details of the analysis in an appendix and some in the text. A second example is *Preparing for Life: early childhood intervention* (PFL 2013) which is a part of an ongoing evaluation of an experimental programme in a disadvantaged community in Dublin which is accessible to the non-academic reader but also has ample statistical detail. Ultimately for an evaluation to be convincing there has to be sufficient detail for the interested reader to assess the strength and appropriateness of the comparisons being made.

---

**c. The 2014 National Assessments of English reading and mathematics**

Subsequent to the publication of this report, the Education Research Center published the 2014 National Assessments of English Reading and Mathematics (Shiel, Kavanagh & Miller (2014), covering primary schools in Ireland (both DEIS and non-DEIS). These assess the changes between May 2014 and the previous national assessments in 2009, a period which partly overlaps with the period considered by Wier & Denner (2013). Clearly this was not available to the latter authors. Nonetheless the results provide an interesting perspective on the Wier & Denner (2013) study, albeit with the benefit of hindsight. These assessments are detailed and comprehensive and I do not claim to have studied them thoroughly. The authors present results for all schools together and for DEIS & non-DEIS schools separately. The first thing to note is that both reading and maths scores rise nationally for both 2\textsuperscript{nd} and 6\textsuperscript{th} year pupils (the two grades studied), with effect sizes around 0.27.\textsuperscript{2} It appears that the increase for schools in DEIS band was of a similar magnitude to the national average (which, of course, includes all DEIS schools). Effects sizes for DEIS band 2 are in some cases much bigger with effect sizes around 0.60.

\textsuperscript{2} An effect size is a standard measure of the effect of a programme. It generally equals the difference in the outcome for the two groups divided by the overall standard deviation although there are variants on this.
Importantly, the authors note (p xvi) “The effect sizes for overall reading suggest that, while substantive improvements have been made in DEIS schools since the 2009 National Assessments, there has been no real reduction in the gap between pupils in DEIS urban schools and in other school types, except at Second class in Band 2 schools”. This would appear to qualify any conclusions one might otherwise draw about the success of the DEIS program.

In principle, the data used for this study is far better suited to make a comparison between DEIS and non-DEIS schools. However I am unable to discern whether this latter study explicitly tested whether the differences between the three groups (DEIS 1 & 2, non-DEIS) were statistically significant. The tests reported appear to refer only to differences between 2009 and 2014, as far as I can see. However as noted on page 8, a simple comparison between the averages of the two groups is likely to underestimate the effect of the assignment to DEIS since we are not comparing like with like.

**Conclusions**

As discussed above, the absence of a control or comparison group severely limits the extent to which this evaluation tells us what we need to know: whether the programme helps the participants, and, if so, the extent of that impact. It is not safe to conclude therefore, on the basis of the evidence presented, that the DEIS programme is working - and as the DES’s press release of December 13th 2013 claims.

What is particularly difficult to understand here is that lack of an appropriate control was apparently built into the design of the evaluation. In other words, this is not a case where the researchers were tasked to carry out an evaluation after the event and where they had no control over what data was collected. I also note the striking lack of sufficient statistical detail in the evaluation.

One further issue that is not discussed is whether the outcomes measured here correspond to what the programme attempts to achieve. That is, is the SSP just about literacy and numeracy or are there other outcomes targeted? The government’s Action Plan for DEIS (DES 2005, section 3.4) refers to “measures to enhance attendance, educational progression, retention and attainment”.

11
The role of the commissioning organization (for example a government department) is critical to the success of an evaluation\(^3\). Not only does the organization fund it but it is often involved in some oversight role in the course of the evaluation and will often have some influence over the nature of the evaluation design. So it is essential that it allows, or indeed requires, the researchers to follow best practice. For example, some organizations are reluctant to fund the collection of data on a control group whether from statistical naïveté or cost considerations. If it is the latter, it is the falsest of economies - it is imperative that this does not happen. In addition, while the commissioning organization cannot be expected to have the same specialist knowledge as the researchers, it must have sufficient knowledge to know whether the evaluation design is appropriate and to know, on completion, how to correctly interpret the results.

Ultimately, it is the role of the commissioning organization to insist that the evaluation answers the question: *does the programme work?* In this context I welcome the establishment of the *Irish Government Economic and Evaluation Service* (www.igees.gov.ie) which aims to strengthen these capabilities in civil service departments.

A good model for policy makers and evaluators is the United Kingdom’s *Educational Endowment Foundation* which was established to fund, develop and evaluate cost effective and replicable projects to eliminate educational disadvantage based on the rigorous evaluations of trials.

\(^3\) I emphasize that these reflections are general and are not a comment on any of the organizations or individuals involved.
Bibliography


