Pre-testing in EEF evaluations

31.10.2013

Education Endowment Foundation

This document discusses the advantages and disadvantages of conducting a pre-test. It looks at the added value of collecting immediate pre-test data over using data in the National Pupil Database (NPD). Baseline covariates should always be controlled for as they significantly improve the precision of estimates. Because KS1 and KS2 performance correlates highly with later performance in Key Stage tests little added precision is provided by conducting a pre-test. Therefore pre-tests should only be conducted if they are needed to evidence the causal model for an intervention. We welcome feedback on this paper.

BACKGROUND

The EEF ultimately wants to demonstrate the impact of its projects on children's attainment at school, in particular at the end of Key Stage 2 at 11 and GCSE results at 16, and can do this by tracking pupils using the National Pupil Database (NPD). If Key Stage 2 or GCSEs do not occur immediately after an intervention then standardised tests that are predictive of national performance are used to measure the immediate impact of the project on children's attainment.

Because most EEF projects are being evaluated using randomised controlled trials, the main outcome will be a comparison of the attainment of the treatment group with that of the control group at the post-test. For this reason, a pre-test is not absolutely necessary. However, pre-tests can be useful for a number of reasons, including demonstrating that that the samples are balanced at the baseline and improving the precision of estimates. But testing is costly and can be burdensome for schools which leads to attrition. This paper sets out the trade-offs involved in a decision about whether to use a pre-test or NPD data. It proposes that in most cases NPD data is sufficient and little added precision is gained by conducting an immediate pre-test.

PROS AND CONS OF USING A PRE-TEST

This section outlines the advantages and disadvantages of conducting a pre-test and some potential alternative solutions. Ultimately, a pre-test should only be used if the advantages outweigh the costs and that decision should be made jointly by the EEF, evaluator and project delivery team.

Advantages

The advantages of conducting a pre-test include:

- **Determining whether samples are balanced at the baseline:** If you randomise a sufficiently large number of units (e.g. pupil or schools) they should be balanced at the baseline in terms of all characteristics. However, there is a risk, particularly with small samples, that balance is not achieved. Also, participants tend to drop out of studies, more often from the control group. Having a pre-test allows you to determine whether the final sample is balanced at the baseline and whether any drop-out has biased the outcome.

- **Improves the precision and power of your estimate:** Even if the samples are balanced at the baseline, using something like ANCOVA where the post-test effect is adjusted using a pre-test covariate improves the precision and power of the analysis. We have provided guidelines to evaluators that prior attainment should be controlled for where possible, but not using gains scores. The more highly correlated the pre-test is with the post-test, then the greater the precision.

- **Understanding the causal model:** Having a pre-test that has various subscales, such as a reading measure that includes word recognition, decoding and comprehension, enables evaluators to look at how an overall effect is moderated by aspects of prior attainment. This

helps to illuminate the causal pathway for how the intervention improves overall attainment. This is most important for testing early stage interventions that are being refined and defined, and understanding how best they should be targeted.

- **Demonstrate the commitment of schools to be involved in the study:** A final advantage of having a pre-test is that it is a way of determining how serious a school is about being involved in an evaluation. For this reason randomisation should always be done after the pre-test so that any schools that drop out of this data collection and not included in the final sample.

Disadvantages

Disadvantages of conducting a pre-test:

- **Costly:** Testing is one of the largest evaluation costs. It is particularly expensive when testing very young children as it has to be done individually or in small groups.

- **Can delay the start of the intervention:** Testing takes time and usually evaluators will give schools a window of one month to complete the tests. If the pre-test is being conducted at the start of the school year, this means that the intervention cannot start until the pre-testing is completed. This can mean that the schools have less time implementing the intervention meaning that there is less of a change of finding an effect.

- **Alternative school or prior pupil covariates could be sufficient:** The trade-off between having an immediate highly correlated pre-test or using alternative pupil level data in the NPD is unclear, and additional precision provided may be relatively small.

The key trade-off we need to understand is exactly how much improvement there is in having highly correlated immediate pre-tests compared to using data that is in the NPD and may have been collected a year or more prior to the intervention and therefore less highly correlated with the post-test. The next section deals with this issue.

USING COVARIATES TO IMPROVE PRECISION

There is always an alternative to conducting a pre-test in English schools, because we have the NPD which contains prior attainment data on all pupils. The table below shows the points at which there is data collected at a national level that could be used as a pre or post-test. The EYFSP, KS1 and KS2 assessments can all be used as a pre-test (or 'baseline covariate'). However, you would expect KS1 will be a better pre-test for Year 3 than it would be for Year 6 because less time will have elapsed and pupil performance should be more highly correlated. However, it is important to understand how much the correlation changes to understand whether it is worth paying for more recent tests.

Table 1: Year groups and national test data in the NPD

Year groups and national test data in the NPD	
Primary	Reception
	Early Years Foundation Stage Profile (EYFSP)
	Year 1
	Year 2
	KS1 teacher assessment (English and Maths)
	Year 3
	Year 4
	Year 5
	Year 6
	KS2 standardised test (English and Maths)
Secondary	Year 7
	Year 8
	Year 9
	Year 10
	Year 11
	GCSEs

Literature from the US on using covariates

Howard Bloom and colleagues in the U.S. published a paper in the Journal of Educational Evaluation and Policy Analysis (2007) that looked at this issue.[1] They used data from five school districts in the U.S. to look at how much controlling for baseline covariates improves the precision of studies that randomise schools to measure the impact of interventions on attainment. They found the following key findings:

- Controlling for baseline attainment can reduce the number of schools required for a given level of precision by half for primary schools and one tenth for secondary schools.
- School-level baseline covariates are as effective as pupil level (although this is not relevant for EEF as we always have access to pupil level information in the NPD).
- The additional precision declines only slightly as the number of years between the pre and post-test increases.
- The added precision is substantial even when the pre-test differs from the post-test.

The results are shown in Table 2 which shows the minimum detectable effect for a cluster randomised trial of reading in three different age groups, with and without a baseline.

[1] Bloom, H., Richburg-Hayes, L. and Black, A.R. (2007) *Using Covariates to Improve Precision for Studies that Randomise Schools to Evaluate Educational Interventions*. Educational Evaluation and Policy Analysis, 29, No.1, pp.30-59.

Table 2: Minimum detectable effect size for a 40 school randomised trial on reading

Minimum ES	Grade 3 (Year 4)	Grade 5 (Year 6)	Grade 10 (Year 11)
No covariate	0.39	0.38	0.42
Covariate 1yr prior (school)	0.26	0.27	0.11
Covariate 1yr prior (pupil)	0.26	0.26	0.10

Using covariates in England – NPD or standardised tests?

An equivalent paper on using pupil or school level covariates in the UK has not been written, although Carole and David Torgerson in their Handbook on conducting randomised trials in education say that having a pre-test with a strong correlation of 0.7 or more can half the required sample size.[2].

In this section data on correlations between English Key Stage tests and standardised tests of attainment are used to understand how much added value is gained by having an immediate pre-test. For example, we are interested in how much added precision is provided in the two following scenarios:

1) Evaluations in Year 6 that use KS2 as the outcome: what is the added value of conducting an end of Year 5 pre-test over using KS1?
2) Evaluations in Year 10 and 11: what is the added value of using an end of Year 9 pre-test over KS2?

Table 3 shows the correlation coefficients for different Key Stages tests and three commonly used standardised tests of attainment developed by CEM Centre at Durham University (PIPs in Years 4 and 5 and Insight in Year 9) and GL Assessments (Year 7 Cognitive Abilities Tests).

These figures show that the PIPs tests in Years 4 and 5 are not much more highly correlated with KS2 performance than the KS1 teacher assessments. Similarly the CAT and Year 9 Insight tests are not much more highly correlated than KS2 is with GCSE performance. The most predictive test is the Year 9 Insight Curriculum test but the gains are still small. The English Insight test is correlated 0.76 with English GCSE compared to 0.69 for KS2 English, and the Maths Insight curriculum test is correlated 0.82 with Maths GCSE, compared to 0.76 for KS2 Maths.

Table 3: Correlations between baseline covariates and post-tests for different outcomes[3]

Baseline covariate	English outcome		Maths outcome	
	KS2	**GCSEs**	**KS2**	**GCSEs**
KS1	0.73	-	0.77	-
Year 4 test (PIPs)	0.759	-	0.765	-
Year 5 test (PIPs)	0.776	-	0.776	-
KS2	-	0.69	-	0.76
Year 7 CAT test		0.66		0.74
Year 9 test (Insight curriculum)	-	0.76	-	0.82
Year 9 test (Insight ability)	-	0.72	-	0.79

[2] Carole Torgerson and David Torgerson (2013) *Randomised trials in education: An introductory handbook.* EEF.

[3] These are all using fine grades (average points scores) and mean GCSE grade as opposed to capped points score as the correlations are higher in these cases. Correlations for the PIPs and Insight are provided by CEM. KS1, KS2 and CATs correlations are provided by FFT: http://www.fft.org.uk/News/FFTBlog/September-2013-(1)/An-analysis-of-Key-Stage-2-reliability-and-validit.aspx#.UnoarnmYaUk).

Box 1: Calculating the Minimum Detectable Effect when using a baseline covariate

The following formula is identified for estimating the minimum detectable effect when using a baseline covariate (Bloom et al.):

$$MDES = M_{J-K}\sqrt{\frac{\rho(1-R_2^2)}{P(1-P)J} + \frac{(1-\rho)(1-R_1^2)}{P(1-P)nJ}}$$

- MDES = minimum detectable effect size
- M_{J-K} = a degrees-of-freedom multiplier1
- J = the total number of schools randomized
- n = the number of students in a grade per school
- P = the proportion of schools randomized to treatment
- p = the unconditional intraclass correlation (without a covariate)
- R_1^2 = the proportion of variance across individuals within schools (at level 1) predicted by the covariate
- R_2^2 = the proportion of variance across schools (at level 2) predicted by the covariate

(NB. For 20 or more degrees of freedom MJ-K equals 2.8 for a two-tail test and 2.5 for a one-tail test with statistical power of 0.80 and statistical significance of 0.05)

We use the formula in Box 1 and Optimal Design Software to calculate the minimum detectable effect size for a 40 school cluster randomised trial with cluster size of 30 and an intra-cluster correlation of 0.2. The effect sizes are given in Table 4 below.

Table 4: Minimum detectable effect size for different baseline covariates and outcomes for a 40 school randomised trial with 30 pupils per cluster, an ICC of 0.2, 0.05 significance, 0.8 power[4]

Baseline covariate	English outcome		Maths outcome	
	KS2	GCSEs	KS2	GCSEs
None	0.43	0.43	0.43	0.43
KS1	0.32	-	0.30	-
Year 4 test (PIPs)	0.30	-	0.30	-
Year 5 test (PIPs)	0.30	-	0.30	-
KS2	-	0.33	-	0.30
Year 7 CAT test	-	0.34	-	0.31
Year 9 test (Insight curriculum)	-	0.30	-	0.28
Year 9 test (Insight ability)	-	0.32	-	0.29

This table shows that the added value of conducting an immediate pre-test at the end of Year 5 or Year 9 over using KS1 or KS2 data is minimal. At most the minimum detectable effect reduces by 0.03 from using an immediate pre-test (using Insight English Curriculum test over KS2 English performance). The gradient of the power curve for a cluster trial with a smaller number of clusters (e.g. 20) is the same, so there is little argument for using an immediate pre-test over Key Stage data to improve power.

[4] These MDES were calculated using Optimal Design Software: http://www.wtgrantfoundation.org/resources/consultation-service-and-optimal-design

GENERAL PRINCIPLES

These findings would indicate that for EEF projects the following principles should be applied:

1. **Evaluators should *always* include a baseline covariate** in the analysis. This will improve precision even when the two samples are balanced at the baseline.

2. **Evaluators should use NPD data as the baseline covariate where possible.** Paying for a pre-test that is the same or similar to the post-test immediately before the intervention is unlikely to add a huge amount of additional precision than using baseline data in the NPD.

3. **A pre-test should only be conducted if needed to understand the causal model.** The only good reason for conducting an immediate pre-test, instead of using NPD data, is if data is needed on precise aspects of pupils' attainment in order to understand and refine the causal model for an intervention. This should only be the case in relatively early stage, developmental evaluations.

4. **If you do conduct a pre-test** e**nsure it is conducted prior to randomisation:** If you do decide to conduct a standardised attainment test as your pre-test and are conducting a randomised trial, make sure that the data is collected prior to randomisation. This is because it demonstrates commitment from the schools, and because there may be drop-out from the control after allocation is revealed and the pre-test will allow you to determine whether the samples are still equivalent at the baseline.