REACH

IFS and IPSOS MORI
Elaine Kelly

Evaluation Summery	
Age range	Year 7
Number of pupils	c.486
Number of schools	27
Design	Randomised controlled trial, with three arms, randomised at the pupil level.
Primary Outcome	Reading

Evaluation of REACH

1. Introduction

The Institute for Fiscal Studies (IFS) and Ipsos MORI are pleased to be able to tender for the evaluation of the University College London Reading Intervention and Reading for Meaning (RIRM) intervention. We bring a wealth of experience in the evaluation of education interventions from both quantitative and qualitative perspectives, including specific experience of the evaluation of a range of literacy interventions, such as the Every Child a Reader programme on behalf of the Department for Education, as well as the LIT and ICAN evaluations on behalf of the Education Endowment Foundation (EEF). We have also joined forces on a number of previous evaluations, including the evaluation of the Saving Gateway on behalf of HM Treasury and have recently been awarded the evaluation of the Achieve Together programme on behalf of EEF.

2. Project team

The evaluation will be led by **Elaine Kelly**, a Research Economist in the Skills sector at IFS. Elaine will be the main point of contact for the project team and EEF, and will take overall responsibility for the quantitative impact analysis. Elaine is currently leading a large-scale programme of research on healthcare on behalf of the Nuffield Trust. She has also worked on a number of projects funded by the Department for Education (DfE), including ongoing analysis of the determinants of risky behaviours, such as teenage pregnancy. Elaine will ensure that the evaluation as a whole is delivered on time, within budget and to a high standard.

Claire Crawford, Programme Director of the Skills sector at IFS, will provide managerial oversight and quality assurance. Claire brings a wealth of experience of education evaluations, including several on behalf of DfE and EEF. In particular, Claire and Elaine are working together on the ongoing ICAN evaluation, so will provide continuity in terms of developing a relationship with the project team and EEF.

The Ipsos MORI team will be led by **Annabelle Phillips**, Research Director, who will provide intellectual leadership for the process evaluation. Annabelle has over 15 years research experience and primarily focuses on qualitative research with young people from deprived areas on subjects such as education, skills, careers and aspirations. She has also directed several evaluations, and is currently directing an evaluation of Jobs Growth Wales for the Welsh Government, and Achieve Together for EEF with IFS. Annabelle will have overall responsibility for the process evaluation, ensuring it runs to time and budget. She will be supported by **Fay Yorath**, Research Manager, who will manage the process evaluation on a day-to-day basis and will be the main point of contact within Ipsos MORI. Fay has worked for a range of public sector clients including DfE, Ofqual, the Home Office, the Youth Justice Board and the Joseph Rowntree Foundation. She is also managing the process evaluation phase of the Achieve Together evaluation. Fay regularly works on qualitative and quantitative projects focused on education, child and young person well-being and development.

3. Overview of the intervention and the role of the independent evaluator

The RIRM intervention is designed to improve the reading skills of those with reading difficulties in Years 7 and 8. Two treatment models are being tested: the first is a 20 week reading intervention programme; the second is a 20 week reading intervention programme plus language comprehension training. Both interventions will target pupils who demonstrate poor reading skills and will involve trained TAs working one-to-one with selected pupils in the summer term of Year 7 and the autumn term of Year 8.[1]

The proposal suggests that 18 children with the lowest reading skills will be selected in each of 27 schools (i.e. 486 pupils in total), and that these pupils will be randomly allocated to one of the two treatment groups or to a wait-listed control group (6 pupils per group)[2]. The wait-listed control group will receive the intervention that proved most successful after the initial 20 week treatment period is over (running in the spring and summer terms of Year 8). The 18 pupils that take part will be selected on the basis of a single world reading test (SWRT[3]) administered to all those who scored below level 4 at Key Stage 2.

Two tests will be used to assess the effectiveness of the intervention: the New Group Reading Test (NGRT; primary outcome) and the full York Assessment of Reading Comprehension test (YARC, of which the SWRT is one component; secondary outcome). Both tests will be administered to all pupils in the treatment and control groups immediately before and at the end of the initial treatment period (i.e. in March and December 2013). The YARC will additionally be administered half-way through the initial treatment period (i.e. in July 2013) and again in April and July 2014, at the mid and end points of the intervention for the wait-listed control group.

[1] Because the treatment groups will receive the interventions spread across the summer term of Year 7 and the autumn term of Year 8 – i.e. over a period separated by the long summer break – it is possible that they may suffer from "summer learning loss". In this case, the overall effectiveness of the interventions may be under-estimated, as TAs may have to spend time repeating material that was covered in the initial 10 week period but subsequently forgotten. We will aim to pick up any concerns that TAs/schools may have about this as part of our light-touch process evaluation (see discussion below for further details).

[2] This is a slight change from the original proposal, which proposed 24 pupils in each of 20 sample schools.

[3] Part of the York Assessment of Reading Comprehension test.

With this information, it will be possible to produce various different impact estimates of the programme on both the absolute level of children's reading skills and their relative progress, as follows:

- the effectiveness of the Reading Intervention relative to the control group after 20 weeks (and 10 weeks using the YARC);
- the effectiveness of the Reading Intervention plus language comprehension training relative to the control group after 20 weeks (and 10 weeks using the YARC);
- the effectiveness of the Reading Intervention relative to the Reading Intervention plus language comprehension training after 20 weeks (and 10 weeks using the YARC);
- to what extent the effects of the programmes (as measured by the YARC) fade in the 10 and 20 weeks after completion[4];
- to what extent the timing and age of intervention matters.[5]

The main aim of our evaluation will be to provide an independent assessment of the quantitative impact of the programme on children's reading skills and to quality assure the results of the project team's evaluation. We are also offering a light-touch process evaluation, which will enable us to better understand how the programme has been implemented and to highlight any issues that might be relevant in terms of potential future roll-out. We will also provide an assessment of the programme's cost-effectiveness.

4. Quantitative impact evaluation

As with all randomised control trials, assuming that the treatment and control groups are well-balanced (i.e. statistically indistinguishable and, given the small sample sizes involved, quantitatively similar) at baseline, it should be possible to obtain unbiased estimates of the impact of the programme by simply comparing the average outcomes of pupils in the treatment and control groups after the intervention.

This assumption highlights the importance of ensuring that the randomisation of pupils is carried out successfully, and we see the independent evaluator playing a key role in this process. IFS would take responsibility for ensuring that pupils in the control group are as similar as possible (in terms of their observable characteristics) to pupils in each treatment group (who would also therefore be as similar as possible to one another). We have extensive experience of selecting comparison groups of pupils and schools for large-scale evaluations (e.g. as part of the FSM evaluation), and have recently randomised the treatment status of schools for the LIT intervention funded by EEF to an extremely tight timescale.

[4] While it will be possible to compare the relative progress made by the two treatment groups after the end of the programme, it will not be possible to provide an estimate of the *impact* of the programmes after they have ended, because the control group – against whose outcomes we would like to compare them – will have started receiving the treatment.

[5] This can be discerned by comparing the progress made over the course of the intervention by the relevant treatment group – who will have received the intervention in the summer term of Year 7 and the autumn term of Year 8 – with that made by the wait-listed control group, who will receive it in the spring and summer terms of Year 8

Due to the tight timescales involved in getting the intervention up and running, NPD data will not be available before the programme has started, hence we will randomise on the basis of gender and baseline test performance only (as has been done for the project team's previous evaluations of similar interventions). We should, however, have access to NPD data for selected pupils at the time of the post test, thus we will be able to use a fuller range of background characteristics (e.g. gender, ethnicity, FSM eligibility, school characteristics, etc) in addition to the pre-test information on reading skills to assess how similar pupils across the three groups are to one another. If it is not possible to perfectly balance the characteristics of these groups of pupils across this larger number of dimensions, then we would use the available characteristics to derive propensity weights for the comparison pupils in order to ensure that the distribution of characteristics amongst comparison pupils matches the distribution of characteristics amongst treatment pupils as closely as possible at baseline.

Given the relatively small sample sizes involved, together with the need to maximise the variation in test scores that we can explain (for the purposes of the power calculations), we will also control for individual characteristics (including baseline test scores) and school characteristics (or school fixed effects) in the impact analysis as well; this will not only help improve the precision of our estimates, but will also be important if the treatment and control groups are not very well-balanced at baseline.

Sample size and power calculations

The proposal aims to recruit a total of 486 pupils across 27 schools, to be split equally across the three groups (two treatment and one control). This means that there will be 12 treated pupils (6 receiving each intervention) and 6 control pupils per school, equivalent to 162 pupils per group. Table 1 below shows the sample sizes required for 80% test power at a 5% level of significance, depending on the estimated effect size (measured in standard deviations, SDs) and the within-cluster correlation in test scores. It assumes that we can explain 60% of the post-test variation in test scores. The left hand panel illustrates the total sample size required (across both treatment and control groups), while the right hand panel focuses on the comparison between a single treatment and control group (which is the relevant sample size for calculating the effect of the intervention).

Table 1 Power calculations

	Within cluster correlation					
	Total sample size (two treatment and one control group)			**Sample size required to detect impact (one treatment group and one control group)**		
Effect Size	0	0.05	0.1	0	0.05	0.1
0.05 SDs	7,535	N/A[6]	N/A	5,023	N/A	N/A
0.1 SDs	1,884	N/A	N/A	1,256	N/A	N/A
0.15 SDs	837	N/A	N/A	558	N/A	N/A
0.2 SDs	**471**	1,069	N/A	**314**	713	N/A

[6] "N/A" in a particular cell means that the number of clusters (schools) is insufficient for the power required, given the effect size and degree of within-cluster correlation.

0.25 SDs	301	456	1,060	201	304	707
0.3 SDs	209	268	390	140	179	260

The project team's power calculations suggested that their proposed sample would be sufficient to detect an effect size of 0.21 SDs assuming zero within-cluster correlation, which is borne out by our own calculations (highlighted in bold above). If pre-test scores alone are able to explain 60% of the variation in post-test outcomes – as suggested by the project team in their proposal – then it may be plausible for us to explain a substantially larger proportion of the variation through the inclusion of relevant baseline factors, including both individual characteristics and school fixed effects. For example, if we were to assume that we could explain 80% of the variation in test scores, then the proposed sample would be sufficient to detect an effect size of 0.14 SDs. If, on the other hand, the within-cluster correlation was greater than zero, say 3%, then – assuming we can explain 60% of the variation in post-test scores – the proposed sample size would only be sufficient to detect an effect size of 0.23 SDs. There is thus a small risk on the downside.

Value for money

It would be straightforward to combine our impact estimates with information on costs (as estimated or collected by the project team) to estimate the cost-effectiveness or value for money (VfM) of the programme, e.g. £/SD improvement in language scores. This could be used to compare the Reading Intervention's value for money with that of other language interventions, including the LIT and ICAN interventions. It is also a reasonably robust and credible measure of VfM, since it avoids the largest challenges involved in cost-benefit analysis (regarding the monetisation of benefits).

5 Process Evaluation

To provide context for the impact evaluation, it is important to understand how the programme is being implemented within the selected schools. We will need to understand key aspects such as: how schools selected the teaching assistants (TAs) to receive training; the quality of the training; how well the programme is being integrated within everyday school life; the impact it has on the pupils and classes; how the TA manages the programme along with other teaching assistant duties; how the programme is managed over two school years within the school; how easy the running records are to keep up to date and their effectiveness as monitoring tools, etc.

Thus, while a 'light touch' process evaluation is required, we feel that a mixed methods approach at different points in the delivery of the programme would yield the most valuable results. It is also important that we maximise the value of the data already being collected as part of this project, and reduce the burden on TAs and any other teaching staff. Therefore, we recommend:

- Analysis of the training evaluation questionnaires collected by the project team
- Analysis of the TA's Delivery Log
- A post project survey with the TAs
- 6 depth interviews with the TAs delivering the programme
- 6 depth interviews with other staff who have been involved in the delivery of the programme

Analysis of the training evaluation questionnaires: the training will be integral to the successful implementation of the programme, so it will be essential to understand how well it prepares TAs to deliver the material and what potential support and accountability mechanisms there may be to ensure fidelity. Each of the TAs will be given an evaluation questionnaire on completion of the training course. Both Ipsos MORI and IFS will feed into the questionnaire design, with Ipsos MORI analysing the results. Findings will be fed back to EEF and the project team as soon as they are available, with a more detailed discussion providing context for the impact evaluation in the final report.

Analysis of the TA's delivery log: this will provide an understanding of whether and how programme material has been delivered on a day-to-day basis and will thus provide important insights into programme fidelity. It will also offer a level of detail that would not be possible to obtain retrospectively through questionnaires and depth interviews. To help ensure that the logs contain the sort of information required for a process evaluation – such as the lessons pupils were taken out of and whether this presented any particular issues, how engaged the pupils were with the programme, etc – we would propose working with the project team to ensure that TAs are aware of the types of information we are looking for at the start of the programme. To meet this aim, we would suggest drafting a covering note and guidance to send to the TA and/or feeding into the training session. Completed logs will be passed to Ipsos MORI for data processing, analysis and reporting at the end of the programme.

A post project survey with the TAs: this will provide an important overview of the programme and how well it worked from the TAs' perspective. For instance, was it easy to deliver? Were there particular issues they had to overcome? What could have been done better? How well was it received across the school? How well was it received by the pupils? What did they perceive its benefits to be? It will also provide the basis from which we select TAs/schools for in-depth interviews (discussed below). Ipsos MORI will work with IFS, the project team and EEF to draft a questionnaire to be asked of all TAs on completion of the 20 week programme (i.e. in December 2013). At this stage, we envisage that it will be a self-completion paper questionnaire, of a maximum of 8 pages, with questions mainly pre-coded with some option for open-ended responses. To minimise the burden on TAs, we recommend that this is the only post-project questionnaire the TAs are asked to complete. We would also recommend that the questionnaires are sent to us rather than to the project team in order to give TAs more confidence in their ability to provide negative feedback anonymously. Ipsos MORI will process the data, analyse and report on the findings.

6 depth interviews with TAs and other teaching staff: on the basis of the TA questionnaires, we would pick out six case studies to explore further. In each case, we would interview the TA and one other staff member (nominated by the TA), in order to provide a broader perspective on how the programme has worked from a school's perspective. We would ensure that all parties are happy with our recommendations before recruitment, but would envisage investigating a range of scenarios. For example, we might include a TA/ school who found the experience very positive, in order to understand what made it successful, with a view to sharing best practice, as well as a

TA/school who struggled or felt like they needed more support, in order to understand what the sticking points were and how these might be overcome in future.

To allow for case study selection, these would be held in January-February 2014.They will be conducted by telephone, and will each last about 30-40 minutes. We would use a 'discussion guide' for these interviews, which can be tailored to each discussion. They will focus on sharing best practice, identifying sticking points, and making recommendations for improving the processes in future. We will analyse the findings, using fieldnotes and a brainstorming analysis session, and write up a summary.

The bigger picture: we will review the material from all of these data sources to form a rounded view of the implementation and fidelity of the RIRM programme, which will be written up and incorporated into the main evaluation report, with a view to providing context for the impact evaluation findings..

6 Risks to the intervention/evaluation

Potential risks to the success of the intervention and subsequent evaluation and how we would mitigate them are described in more detail in Table 2 below.

7 Budget

We have budgeted for the production of a short evaluation report (less than 50 pages) by December 2014, including a non-technical summary and a more detailed methodological appendix. This would encompass the results from both the impact and process evaluations, with the latter mainly providing context for the former. We are happy to present our results to the EEF and/or the project team if required.

In agreement with EEF, we have also budgeted for the New Group Reading Test to be carried out on all children selected for the programme before and after the initial intervention (i.e. in March and December 2014) in addition to the tests proposed by the project team. The total cost of doing so is estimated at £8,832. This is comprised of a licensing cost of £4,800, and project team administrative costs of £4,032.[7]

[7] The cost of the license for the digital version of the NGRT is £4.80 per administration, including VAT. The total cost is therefore £4.80 per test x 500 children x 2 test occasions = £4,800. The administrative costs are based on 2 test occasions x 2 hours time for each test occasion x 2 TAs at each testing session x 42 schools x £12 hour = £4,032.

Table 2 Risks to the evaluation

Risk	Likelihood	Impact	Mitigation and contingency plan
Imperfect randomisation	Low	Low	We would provide independent oversight of the randomisation process, conducting statistical analyses to ensure that the characteristics of pupils in the treatment and control groups are well-balanced. If it is not possible to balance the characteristics of these groups then we would use a variety of non-experimental techniques during the impact analysis to make it as robust as possible (as outlined above). We have considerable expertise in such methods.
Sample sizes too small to detect significant impacts	Low	High	As long as the assumptions provided by the project team – regarding the proportion of post-test variation explained by the pre-test and the degree of intra-cluster correlation – are correct, the proposed sample should be large enough to detect the magnitude of effect sizes expected by the project team. In the relatively unlikely event that these assumptions do not hold, the risk to the evaluation is significant. We will work with the project team to ensure that these assumptions are as accurate as possible prior to the intervention starting.
Control pupils receive some form of treatment as a result of spillovers	Low	Medium	The fact that the intervention is being randomised within school means that there is a small chance of spillover from the treatment group to the control group (e.g. the support provided to control group pupils improves because there are fewer pupils in their class while the treatment group pupils are being treated). If this were to happen then our evaluation would underestimate the overall impact of the programme on children's reading skills. We are reassured by the project team's suggestion that at most two pupils will be taken out of any particular class at any given time.
Difficulty in recruiting schools, teachers and stakeholders for the process evaluation and keeping them engaged	Medium	Medium	The qualitative research team have considerable experience of working with teaching staff and other professionals and we would draw on this expertise to ensure that the evaluation agenda is relevant and useful for participants so as to keep them interested and engaged. We will seek to minimise the burden on participants by ensuring they are recruited early and given clear information in advance about what we expect from them. We will also give them a choice about the best way to interview them and will work flexibly around their time commitments.
Delay in providing data to the evaluation team	Low	Medium	The proposed timetable should allow sufficient time for the project team to collect the data, clean it and pass it on to the evaluation team in a timely fashion. The main risk to the timetable is likely to be following the final post-test in July 2014. We will prepare our code in advance to

			minimise the risk that the final report will not be delivered on time.
Unavailability of staff during project	Low	Low	IFS and Ipsos-MORI each have a pool of researchers on hand with expertise in programme evaluation, education policy and qualitative research methods respectively who would be capable of taking over work on this project. Systems and procedures will be adequately documented to ensure handover can occur smoothly if necessary.
Loss of or damage to data	Low	High	Both Ipsos-MORI and IFS have high levels of IT security in place. All members of the evaluation team have extensive experience of working with data, and are well aware of the importance of keeping data safe and of using the necessary security procedures. Back-ups are located off-site and can be retrieved within one working day. EEF and the project team will be immediately notified if data is accidentally damaged or stolen; contingency plans specified in our security policies (available on request) can then be put into place.

Timeline and reporting

Task	**Month/Year**						
	11/12 to 12/12	01/13 to 03/13	04/13 to 08/13	09/13 to 12/13	01/14 to 03/14	04/14 to 08/14	08/14 to 12/14
Assist project team with selection of schools if required	■						
Randomise pupils to treatment and control groups and assess balancing		■					
Conduct process evaluation			■	■	■		
Quantitative impact analysis carried out							■

Deliver final report (December 2014)							▨
	11/12 to 12/12	01/13 to 03/13	04/13 to 08/13	09/13 to 12/13	01/14 to 03/14	04/14 to 08/14	08/14 to 12/14

Note: these timings assume that the recruitment of schools, carrying out of the pre- and post-tests and delivery of the data to the evaluation team are all done according to the project team’s timetable.