Department for Education

The relative effectiveness of blended versus face-to-face adult English and maths learning

Research report

February 2018

Jake Anders, Richard Dorsett and Lucy Stokes – National Institute of Economic and Social Research

Contents

List of figures

List of tables

Summary

Improving English and maths skills amongst those failing to reach basic standards has been an important focus for government policy for some decades, and continues to be so today. Existing studies have identified adult literacy skills in England as being fairly average by international standards, with numeracy skills ranking towards the lower end of comparison tables. However, it is not the case that skills are low across the population, instead, there is a considerable disparity between the highest and lowest performers, with a sizeable group of adults who have particularly low skills. The Skills for Life strategy aims to prioritise support for groups of adults with the greatest numeracy and literacy need.

In March 2013, the Department for Business, Innovation and Skills (BIS) commissioned a programme of research into adult English and maths. This involved two elements:

- A randomised control trial (RCT) of the relative effectiveness of face-to-face compared to blended learning; that is, learning that makes significant pedagogic use of information and communications technology (ICT); and,
- A longitudinal survey of learners, allowing the dynamics of skills gain and loss to be observed.

This report is concerned with the RCT. The primary aim of the RCT is to assess the impact of delivery mode on skills. The group of learners considered are those aged 19 or over who were enrolling on English or maths courses at any level from Entry Level 1 up to Level 2.

This report documents the RCT design decisions, describes the details of implementation and provides an assessment of how well the trial has worked. As will be seen, the number of learners in the final analysis sample was much smaller than planned, raising concerns about the ability of the trial to detect effects of the size expected. The estimation results are presented along with a discussion of how they should be viewed. The key lessons learned from the project are also summarised, with the aim of informing future RCTs in the sector.

Design of the RCT

The RCT had two treatment arms: “traditional” face-to-face tuition and blended learning. Face-to-face learning is interpreted as using technology for less than 5 per cent of guided learning hours, while blended learning uses technology for between 30 and 60 per cent of guided learning hours. Furthermore, there is a difference in the nature of technology use. In the face-to-face case, learning remains fundamentally teacher-led; in the blended case, technology is used as a complementary part of the pedagogy.

The RCT was conducted by randomly assigning individuals to either a face-to-face learning class or a blended learning class. Individuals were randomised at the point of enrolling for their course. As part of the randomisation process, background information was collected and learners' consent to link their data to other sources, such as the BIS Individualised Learner Record (ILR) and the DWP Work and Pensions Longitudinal Study (WPLS) was sought.

A target of 750 English learners and 750 maths learners was set. It was estimated that this would be sufficient to detect an effect size of 0.17. The RCT provides an estimate of the effect of intention to treat (ITT).

Implementation of the RCT

The first step in recruiting providers was to carry out a survey asking about the nature of their provision. Providers that showed interest in being involved in the trial were followed up in order to establish their suitability and, if appropriate, to obtain their agreement to full participation. To be suitable, they had to: deliver both face-to-face and blended learning; have sufficient numbers of learners for English and maths courses; and be willing to introduce randomisation into their enrolment process. It is likely that this prevented smaller centres from taking part, raising questions about the representativeness of the eventual impact estimates.

It was not until just before the end of the 2012/13 academic year that providers could first be approached. This allowed very little time to introduce randomisation procedures for the next academic year's intake. Consequently, only a handful of providers were in a position to participate in Autumn 2013. Since the numbers achieved in 2013/14 (Phase 1) were low, the intake period was extended to include the 2014/15 academic year (Phase 2). Furthermore, participation was incentivised in Phase 2 through a payment of £500 and, in order to encourage providers to remain with the RCT, two later payments of £10 per learner. In addition, further support was provided during the assessments.

Consent was also sought from learners. They had to be willing and able to attend either the face-to-face learning class or the blended learning class. Those who gave consent were then led through the randomisation process, with their assignment outcome determined according to a pre-defined randomisation sequence. In Phase 1, randomisation was overseen by project caseworkers in order to minimise the burden on centres and to ensure that randomisation proceeded as intended. In Phase 2, randomisation was automated and was trigged when learners submitted the online background questionnaire.

Sample descriptives, randomisation performance and programme fidelity

In total, 13 centres participated in the study, including one large centre that accounted for 42 per cent of all eligible learners. The number of learners randomised was 863, made up of 472 English learners and 391 maths learners. This was below the target number but there was further loss of sample due to only a small proportion of learners being assessed at the start and end of their course. Both assessments are required to give a measure of skill change, the primary outcome of interest. Reading skill was assessed at both points for 74 learners and writing skill was assessed at both points for 58 learners. Maths skill was assessed at both points for 75 learners.

A comparison of blended and face-to-face learners' background characteristics found no significant differences, suggesting that the randomisation had been successful in achieving two broadly similar-looking groups. This is true for eligible learners as a whole and for the subset of learners for whom both assessments were available. Importantly, skills measured at the start of the course were also similar across treatment arms. Furthermore, there is evidence that there was a meaningful difference across treatment arms in the delivery of learning, with the blended learning arm making greater use of ICT. However, with such small numbers of observations, the ability of the trial to detect effects is greatly reduced.

Estimated impacts

For both English outcomes – reading and writing – the estimated effect size associated with attending a blended learning class rather than a traditional face-to-face class was 0.22. That is, blended learning is estimated to increase English skills by 0.22 standard deviations. However, this was not statistically significant. In order to provide some context to help interpret this finding, the results also show the minimum detectable effect size (MDES); the smallest effect size one could expect to detect given the achieved sample. This was 0.29 for reading and 0.31 for writing. So, rather than the non-significant result suggesting that mode of learning is unimportant, it should be taken to mean that it did not have a big enough effect to register. For maths, the estimated effect size was 0.10 and again this was not significant. The MDES in this case was 0.25 so again, the appropriate interpretation is that the effect was smaller than 0.25 standard deviations.

Conclusion

The primary aim of this RCT was to test whether blended learning and face-to-face learning differ in their effectiveness at increasing skills. This was a complex study

involving a number of significant practical challenges, from recruiting colleges to administering skills assessments. In the event, despite the randomisation and implementation going well, the small number of assessments reduced the statistical power of the RCT to the extent that the results are rather inconclusive.

While this is disappointing, there are a number of key learning points from this study:

- The RCT was resource intensive for providers. They lacked the capacity to take on the additional work that the RCT presented, especially when facing many competing demands. Providers required substantial external support, and would have required considerable financial incentives to enable their existing staff to have the necessary time to take part in the research. However, even when offered, not all providers took up the offer of additional help.
- Designing the research in order to minimise the additional burden on providers is key. This includes careful consideration of requirements for data collection, which in this case proved burdensome. The assessments were also lengthy, taking up substantial and valuable class time; it is important to strike a balance between assessments that are robust and fit for purpose against the practicalities of administration.
- Many providers were not ready to deliver blended learning to the extent that the RCT had envisaged – not all providers, for example, had the necessary infrastructure and technology in place. The timeline for the evaluation was very ambitious and putting in place the necessary arrangements to deliver the trial from the start of the 2013/14 academic year proved too great a challenge for many providers.
- As far as is feasible, the RCT needs to fit alongside the practicalities of the day-to-day running of provision. In this study for example, there was a tension between the need to randomise learners to the different modes of learning and offering learners flexibility and choice.
- To be a success, an RCT requires commitment from all parties involved. It is encouraging that senior leaders were enthusiastic about participating in the research. However, in some cases teachers and learners needed reassurance about the motivation behind the study as well as other forms of support, such as adequate time for training in blended learning for teachers.

Many of these lessons are applicable to the design and delivery of RCTs both within and beyond the FE sector. It is also relevant to point out that the experience from this study is not sufficient to draw the general conclusion that RCTs have no potential role to play in building the skills and training evidence base. However, it does demonstrate how important practical considerations can be.

1 Introduction

1.1 Background

There are some skills that are fundamental: to be successful in life and at work, people must be able to read and write and to use numbers with confidence. People need these skills for a functioning society and a healthy economy.

Improving English and maths skills amongst those failing to reach basic standards has been an important focus for government policy for some decades, and continues to be so today, indicated by the Government's statutory entitlement for adults to access fully-funded English and maths courses to progress to GCSE-Level 2.[1,2]

Existing studies have identified adult literacy skills in England as being fairly average by international standards, with numeracy skills ranking towards the lower end of comparison tables.[3] Notably, international studies both of adults and of young people (aged 15)[4] reveal a population that is characterised by a comparatively wide distribution of skills. The issue is not that skills are low across the population, but that there is a sizeable disparity between the highest and lowest performers. In essence there was, and remains, a large group of adults who have particularly low skills.

Government introduced the Skills for Life strategy in 2001 to address this issue by prioritising groups of adults with the greatest numeracy and literacy need. This includes free education and training provision to enable adults with poor basic skills to develop their literacy and numeracy skills. In terms of learner numbers the strategy has had great success, surpassing its 2004 target of supporting 2.25 million adults to achieve a Skills for Life qualification by 2010, with 2.8 million learners achieving a Skills for Life qualification by 2009.[5] However, the Skills for Life surveys of the general adult population in 2003 and 2011 reveal that while the proportion of adults showing literacy skills of Level 2 standard and above has increased (equivalent to grades A*-C at GCSE),

[1] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/485969/BIS-15-615-skills-funding-letter-2016-to-2017.pdf
[2] The intention to improve the quality of apprenticeships outlined in 'New Challenges, New Chances' also remains a key Government priority (for example see https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/482754/BIS-15-604-english-apprenticeships-our-2020-vision.pdf) however apprenticeships was beyond the scope of this research.
[3] https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/246534/bis-13-1221-international-survey-of-adult-skills-2012.pdf

[5] Department for Innovation, Universities and Skills (2009). Skills for Life: Changing Lives.

there has been a disappointing level of progress at the lower end of the scale.[6] Numeracy skills actually appear to have shown a small decline.

Following consultation, in 2011 the Department for Business Innovation and Skills' New Challenges, New Chances report outlined reform plans for the further education and skills system. This includes expanding the Skills for Life programmes to offer free courses to enable adults to improve their basic literacy and numeracy skills to English and maths GCSE / Level 2.[7]

In 2013, the Department for Business, Innovation and Skills (BIS) commissioned a consortium of research organisations to carry out a programme of research into adult English and maths. This involved two key elements:

- A randomised control trial (RCT) of the relative effectiveness of face-to-face compared to blended learning; that is, learning that makes significant pedagogic use of information and communications technology (ICT)
- A longitudinal survey of learners, allowing the dynamics of skills gain and loss to be observed.

All learners included in the research were aged 19 or above and attended English or maths courses between Entry Level 1 and Level 2. The research was delivered by a consortium of organisations: Kantar Public (formerly TNS BMRB) conducted the longitudinal survey, drawing on assessment tools designed by AlphaPlus, support from NIACE in the recruitment of colleges, and analysis by NIESR. The RCT was led by NIESR and AlphaPlus, with support from NIACE. Additionally, during the development stages Professor Steve Reder at Portland State University offered his expertise into the questionnaire design and analysis.

This report is concerned with the RCT. As stated in its protocol (see Appendix A), the primary aim of the RCT was to assess the impact of delivery mode on skills at the time of course completion. Secondary outcomes were identified as longer-term attainment, confidence with English and maths, labour market outcomes and subjective well-being.

This report documents the RCT design, describes the details of implementation and provides an assessment of how well the trial has worked. Where there have been deviations from the protocol, these are highlighted and discussed. Impact estimates are presented and discussed.

---

[6] BMRB Skills for Life (2003); TNS BMRB Skills for Life (2011).
[7] https://www.gov.uk/government/consultations/new-challenges-new-chances-next-steps-in-implementing-the-further-education-reform-programme

1.2 Existing evidence on effect of mode of learning

Existing evidence is patchy and there are a number of reasons why the results from previous studies are not of direct relevance. First, studies differ in what they regard as blended learning. As discussed in Section 2.1, reaching a clear and workable definition of blended learning requires judgement. While care was taken to draw on expert advice in this study, it is inevitable that other studies will use different definitions. Furthermore, with increasing prevalence of ICT in the classroom, the baseline against which blended learning is assessed has changed. This calls into question the extent to which older studies can be viewed as informative of the relative impact of more modern technology in teaching. Second, there are very few studies that consider the case of adult basic skills learners. More common is to consider university students. Such students are different in fundamental ways from the learners considered in this study, not least in how comfortable they may feel with technology. Consequently, there is no reason to expect the mode of learning to have the same impact as it would on our target group. A third reason is that the available evidence varies in the nature of the outcomes considered. In this study, Item Response Theory (IRT) was used to derive measures of individuals' skills. IRT assumes that individuals have an underlying level of skill and ability and that this influences their observed performance in test questions. Using individuals' responses to numerous test questions IRT provides a framework that allows underlying ability to be estimated. For learners of English, the IRT analysis created skills estimates in two domains: Reading and Writing. For maths learners, there was a single skills measure. Other studies typically use simpler outcomes that do not claim to directly represent skills.

Despite these shortcomings, it is only through existing studies that we can get any insight into the kind of effect size we might expect. Hattie (2009) in his synthesis of learning meta analyses shows many education interventions to have small effect sizes but that computer-assisted instruction (CAI) averaged an effect size of 0.37. This was greater for English than maths (ranging from 0.35-0.73 for various aspects of English but 0.21 for maths). Bernard et al. (2004) show that there is considerable variation across type of learner. While they conclude that the effect of e-learning compared to face-to-face learning is essentially zero, many of the subjects in the studies they consider were undergraduates. The small number that instead considered military personnel – arguably closer to the type of learner in our study – showed a significant effect size of 0.45.

A 2010 U.S. Department of Education meta analysis has the advantage that it considers more recent studies (most were published in 2004 or later), which perhaps better reflect more recent technology. For e-learning, the mean effect size was 0.05 while for blended learning there was an effect size of 0.35. A caveat is that these estimates could not control for the fact that the experiences of students with different modes of learning also differed in other ways, notably the amount of time devoted to learning. Hence, the results may reflect these differences, including the amount of learning, as well as the mode itself.

Taken together, it is clear that there is some uncertainty about what effect size to expect. In view of this, the RCT was designed to maximise its chances of detecting a small effect (that is, maximising statistical power).

1.3 Content of report

The remainder of this report has the following structure. Chapter 2 sets out the detail of the RCT and gives the reasons behind the major design decisions. It also summarises deviations from the trial protocol. Chapter 3 describes implementation and how randomisation worked in practice. The available data are described in Chapter 4. Chapter 5 presents summary statistics based on these data. It considers how well the RCT has worked in terms of achieving two similar-looking groups of learners and in terms of delivering two types of learning that are sufficiently distinct with regard to their use of ICT. The estimation results are presented in Chapter 6. Chapter 7 sets out conclusions, including lessons learned.

2 Design of the RCT

2.1 Defining treatment arms

A prerequisite for the trial was to identify a working definition of blended learning. A concern was that blended learning may be so varied that treating it as a single category would end up with a group of students, some of whom experience something very similar to the face-to-face learners. Blurring the distinction between the categories would make it harder to detect effects, increasing the risk of inconclusive results. To mitigate against this, the two modes of learning were distinguished by the extent and nature of their use of technology[8].

- With regard to the extent of use, we interpret:
  - face-to-face learning as using technology for less than 5 per cent of guided learning hours
  - blended learning as using technology for between 30 and 60 per cent of guided learning hours.
- With regard to the nature of the use of technology:
  - Face-to-face learning may include an element of technology to practise or consolidate skills through self-study but this must not form a significant part of the course. Tutors may use the web for ideas, activities and resources to use in their face-to-face teaching and the class may use computers for one-off sessions but this should be teacher-led.
  - Blended learning combines multiple delivery media that are designed to complement each other and promote learning and application-learned behaviour. This may include several forms of learning tools, such as real-time virtual/ collaboration software, self-paced web-based courses, electronic performance support systems embedded within the job-task environment, and knowledge management systems. Blended learning often may be a mix of traditional instructor-led training, synchronous online conferencing or training, asynchronous self-paced study, and structured on-the-job training from an experienced worker or mentor

There is, of course, still scope within this definition for considerable variation in the nature of blended learning. In particular, providers may vary in the online learning packages they use. However, it is not the aim of this study to evaluate specific ICT tools. Instead, the focus is on understanding whether the use of ICT as it currently stands is related to attainment and skills. There is a need to ensure, as far as possible, that the arms of the trial differ only in the mode of delivery and not in other regards. Particularly important is

[8] These definitions were based on the recommendations of a Delphi panel of education and education technology experts.

that learners in both arms of the trial receive a similar number of learning hours. If this is not the case, the estimated effects will be confounded in the sense that it will not be possible to tell whether they are driven by the mode itself or by the number of hours. With this in mind, providers were expected to offer courses that comprised a similar number of guided learning hours.

2.2 Individual randomisation

The RCT was conducted by randomly assigning individuals to either a face-to-face learning class or a blended learning class. Randomisation took place at the point of course enrolment. During randomisation, baseline information on learners was collected. We use this in Section 5.2 to show both the characteristics of the randomised sample and the extent to which the randomisation appears to have successfully created two similar-looking groups. Learners were also asked at randomisation for consent to link their data to other sources, such as the Individualised Learner Record (ILR) and the DWP Work and Pensions Longitudinal Study (WPLS).

2.3 Power calculations

As discussed, there is little reliable guidance as to the expected effect size. Consequently, the approach taken kept in mind the need to maximise power by including as many learners as possible and by collecting rich background data that could be included in the eventual estimation. The target numbers for the trials were 750 English learners and 750 maths learners. We assume that focusing on the change in attainment rather than post-test attainment *per se* will reduce the variance of the outcomes by 30 per cent. Requiring 2-tail tests with 80% power and 95% significance implies, for each trial, a minimum detectable effect size of 0.17.

2.4 Generating randomisation sequences

The randomisation process involved assessing eligibility, collecting background information, requesting consent and then randomising individuals to either a face-to-face learning class or a blended learning class. This last stage was carried out using randomisation sequences that had been previously generated and embedded in the software used to handle the full process. A separate sequence was generated for each learning provider and for each subject (English and maths). Sequences were generated using permuted block randomisation with a 50:50 allocation ratio. In other words, sequences were built up of smaller “blocks”, each of which had four randomisation outcomes: two to the face-to-face learning class and two to the blended learning class. Within each block, the order of these outcomes was randomised. This approach avoids randomisation outcomes being predictable, reducing the scope to circumvent randomisation, while still converging quickly to the required 50:50 allocation.

Furthermore, since the sequences were pre-specified and the precise time of randomisation for each learner was recorded, it is possible to examine whether individuals' randomisation outcomes were correct. This was monitored in the early period of randomisation to ensure that this was being implemented as intended.

2.5 Analytical approach

Effects were estimated separately for English and maths learners, using linear regression to control for observed differences in the composition of treatment and control arms. In principle, linear regression is unnecessary. Randomisation should ensure that the individuals within either treatment arm are similar (on average) at baseline so simply comparing mean outcomes provides an unbiased impact estimate. However, the advantage of regression is that it allows background characteristics to be controlled for, thereby improving the precision of the impact estimates.

These estimates capture the impact of blended learning relative to face-to-face tuition. It should be noted that not all individuals included in the trial will in fact receive the mode of learning to which they are assigned, either because they drop out or because somehow they receive the alternative mode. In keeping with common practice, the trial is designed to capture the effect of being assigned to one arm rather than the other. As such, it provides an estimate of the effect of so-called intention to treat (ITT). With non-compliance of the type described, this can differ from the effect of actually receiving one mode of learning rather than the other.

The main results are based on all learners for whom we have both pre- and post-test outcomes. As described in Chapter 5, the sample available for analysis was considerably smaller than anticipated. In view of this, alongside the estimated impacts and standard errors, we present measures of the minimum detectable effect size in each case. This is relevant to interpreting the estimated impacts; lack of statistical significance should not necessarily be taken to mean that there is no effect, rather it could be that the effect is not sufficiently large to be detected.

2.6 Ethical considerations

The RCT was considered by an ethics group during the design phase. This emphasised the need to enshrine the principle of voluntary informed consent and to make explicit that participants are free to withdraw at any stage. The ethics group also highlighted the need to minimise the effects of designs that advantage one group of participants over others.

In response to the views expressed by the ethics group, a small change was made to the consent question in the background questionnaire administered as part of the

randomisation process. Also, the protocol was altered to mention the risk of learners doing better on one treatment arm than the other.

3 Implementation of the RCT

The nature of the intervention meant that the RCT was heavily reliant on the cooperation of learning providers and learners themselves. In this Chapter, we provide a summary of the recruitment processes. We also describe how the randomisation process operated in practice. As a general comment, the practical details of implementing a RCT are of central importance. While a strong RCT design ensures robust estimates in principle, the extent to which this will be achieved in practice depends on the feasibility of delivering the RCT as intended.

3.1 Recruiting providers

The first step in recruiting providers was to carry out a survey asking about the nature of their provision. Providers that showed interest in being involved in the trial were followed up in order to establish their suitability and, if appropriate, to obtain their agreement to full participation. To be suitable, they had to: deliver both face-to-face and blended learning; have sufficient numbers of learners for basic English and maths courses;[9] and be willing to introduce randomisation into their enrolment process.

It was not until just before the end of the 2012/13 academic year that providers could first be approached. This allowed very little time to introduce randomisation procedures for the next academic year's intake. Consequently, only a handful of providers were in a position to participate in Autumn 2013. Since the numbers achieved in 2013/14 were low (even after including those enrolling in early 2014), the intake period was extended to include the 2014/15 academic year (Phase 2). This resulted in a considerable increase to the sample size, and also meant that the study as a whole included a larger number of centres.

An incentive payment of £500 was introduced to encourage participation in Phase 2 and, in order to encourage remaining with the RCT, providers received a payment of £10 per learner at two points over the course of the study (after the initial £500 payment). In addition, phase 2 colleges could claim up to £10 per hour as a contribution towards a further support person in the classroom during the assessments. Some colleges also requested further support from a project caseworker[10] during busy enrolment periods, which was given.

[9] The initial plan was to include only providers who would offer a minimum of 150 learners to the RCT but some exceptions were agreed.

[10] In Phase 1, each provider had a dedicated caseworker who was responsible for the day-to-day running of the RCT. The caseworker could offer a range of support which included supervising the randomisation and assessment process, as well as CPD support in blended-learning pedagogy. In Phase 2 there were no caseworkers, instead a more devolved approach was adopted with a paid RCT lead in each provider.

Centres also had to be able to deliver both modes of learning in line with the definitions presented above. A series of information events took place with the aim of communicating the requirements of the study so that centres could judge whether they were able to participate.

As a general comment, the recruitment of centres was difficult, perhaps unsurprisingly given the very specific requirements involved. Priority was given to attracting sufficient centres to allow the target number of learners to be recruited. It is likely that the need to be able to offer two modes of delivery for the same subject prevented smaller centres from taking part. This raises questions around the generalisability of the eventual impact estimates (or "external validity"). However, it does not affect internal validity. In other words, the estimated impacts for participants could still be robust.

3.2 Recruiting learners

Only learners within participating centres were eligible to participate in the RCT. In addition to enrolling on English or maths courses at a level between Entry Level 1 up to Level 2, they had to be willing and able to attend either the face-to-face learning class or the blended learning class. Inevitably, this ruled out a number of individuals from participating. To the extent that there is a systematic tendency for particular types of individuals to be excluded from the RCT (perhaps, for instance, those in work are more constrained in the courses they are able to attend), this may affect external validity. As with the recruitment of centres though, internal validity is unaffected.

3.3 Carrying out the randomisation

Randomisation was embedded in the enrolment process. Within each participating centre, the eligibility of those enrolling was first established. Those who were eligible – that is, they were able to attend either a face-to-face learning class or a blended learning class – were then told about the study and asked if they would be happy to participate. Those who were happy then answered a background questionnaire (BQ), designed to take about 10 minutes.

In addition to collecting background information on participants, the BQ also included questions requesting learners' consent to:

- complete the survey
- take part in the research and be randomised
- take part in a follow-up survey[11]

[11] It was intended that the RCT sample would be included in the longitudinal survey sample and so have longer-term outcome information collected. This turned out not to be feasible, since extending the intake

- have their responses given as part of the research linked to other information on adult learning. This, in effect, means the ILR.
- have their responses given as part of the research linked to a learner dataset with benefit and employment data. This, in effect, means the WPLS.

Those consenting to be randomised were then told whether they would be attending a face-to-face learning class or a blended learning class. This was determined by the randomisation sequences described above; the learner simply received the next available randomisation outcome.

Clearly, embedding randomisation into the enrolment process represents a substantial change to providers' practices. A priority throughout was to minimise the extent to which this imposed an administrative or operational burden on provider staff. To achieve this, project caseworkers attended enrolment, overseeing and assisting with the new procedures. It was also important not to burden or inconvenience the learners themselves. The caseworkers explained the RCT to learners and were on hand to help, should those opting to participate have any difficulty completing the BQ, for example. By Phase 2, the process was more automated[12] and randomisation was triggered once learners had completed the BQ.

period for the RCT to include 2014/15 admissions created too much of a delay relative to the timetable of the longitudinal survey.

[12] In Phase 1, it was deemed necessary for someone to be with each learner at the point of randomisation. In Phase 2, following consent given at the end of the online questionnaire, learners were automatically redirected to the randomisation tool and would be allocated to one of the delivery modes, relieving the pressure on the number of staff required at enrolment.

4 Data

The data required for the study were collected through a number of sources. This Chapter briefly describes each of these. In Chapter 5, descriptive statistics based on these data are reported in order to provide an illustration of how the RCT has operated in practice.

4.1 Background questionnaire

As discussed already, the BQ was completed by learners during the enrolment process.[13] It served a number of purposes. First, it collected background information. This covered personal characteristics: age, sex, ethnic group, whether English was the first language, qualifications, household composition and, employment status. It asked about learners' motivation for studying and also included a range of questions assessing the level of difficulty experienced in everyday life as a result of insufficient skills in English or maths (depending on which subject the learner was enrolling for).

This information collected by the BQ helped when estimating impacts, by controlling for some sources of variation within the RCT sample. In addition, there were two other important roles of the BQ. First, it provided a means of obtaining learner consent to participate in the RCT and to link to other sources of data. Second, it collected contact details (names, date of birth, address, postcode, telephone numbers and, email). These, along with information on provider, course and subject (also collected by the BQ), allowed the RCT data to be linked to the ILR and WPLS.

4.2 Randomisation outcome

On completion of the BQ, learners were told which class they were assigned to attend. This was on the basis of the randomisation sequences described earlier. The assignment outcome was saved and merged with the BQ responses.

4.3 Assessment data

Learners' skills were assessed through administered tests. These were designed to take place at the start of their course and again at the end. The test results were analysed using item response models in order to provide an estimate of underlying skills. Item Response Theory (IRT) provides a coherent framework for producing a skills measure that spans different course levels. It assumes that individuals have an underlying level of

[13] There were some minor changes to the background questionnaire in July 2014. Specifically, the consent question was slightly re-worded, in order to make it easier to understand for learners whose first language was not English, and date of birth was asked (rather than age at randomisation), as Phase 1 learners had often included their date of birth when asked for age.

skill and ability and that this influences their observed performance in test questions. Using individuals' responses to numerous test questions IRT provides a framework that allows underlying ability to be estimated. Details of its implementation in this study are provided in Boyle and Horrocks (2015, 2016).

Learners sat the tests in class. For learners of English, the IRT analysis created skills estimates in two domains: Reading and Writing. For maths learners, there was a single skills measure. A comparison of pre- and post-learning test scores provides a measure of skills gain.

4.4 Survey data

The intention had been that individuals participating in the RCT would form part of the sample for the longitudinal survey of learners that formed the other strand of the overall project (see Section 1.1). In the event, this turned out not to be possible since extending the intake period for the RCT to include 2014/15 admissions created too much of a delay relative to the timetable of the longitudinal survey. The reason for wanting to include the RCT participants in the survey data was that this would have allowed impacts on secondary outcomes relating to longer-term attainment, confidence in English and maths and subjective well-being to be estimated.

4.5 Administrative data

It was also intended that the RCT data could be linked to administrative data; the ILR and WPLS. In practice, there were difficulties achieving this so impacts on outcomes taken from administrative data were not estimated. Several factors contributed to this. First, not all RCT participants gave consent to link their data to the ILR and WPLS (see section 5). Second, among those who did consent to ILR linkage, not all could be successfully matched. While individual learners could be matched in more than than 70% of cases, it was not always clear that the learning aims were being successfully matched. Often, the RCT course start dates and subjects did not agree with those recorded in the ILR. More detail is provided in Chapter 5.

4.6 Teacher questionnaires

Teachers in those classes involved in the RCT were given short questionnaires at the start, middle and end of their courses.[14] These questionnaires were designed to collect information on the use of ICT in class. They provide an insight into the nature of the distinction in practice between face-to-face and blended learning classes. The

[14] Where teachers taught more than one class participating in the RCT, they were asked to complete a separate questionnaire for each class.

questionnaire asks teachers to assess the extent to which ICT was used in their class along 13 different dimensions. The first four of these dimensions related to the extent to which ICT was a fundamental part of the teaching and learning, as well as learners' active and passive use of ICT in their learning, while the remaining questions focused on more specific examples of the different uses of ICT in teaching and learning (Murphy, Grant and Smith, 2014). The questionnaire also collected additional information on the class and teacher, including class subject, level, and start and finish dates. Improvements were made to the teacher questionnaire for Phase 2, such that we only report analysis of responses for Phase 2 in this report.[15]

4.7 Monitoring information

Phase 2 centres were asked to provide periodic monitoring information (in Phase 1, this was collected by caseworkers). This provided both information about the classes (teacher name, subject, delivery mode, day and time, class length, guided learning hours, course start date) and information on the learners within each class (name, date of birth, learner start date, learner expected completion date and assessment dates). This can be used, among other things, to estimate the rate of drop out. However, it is worth noting that this information is not complete for all learners, for example, for a substantial proportion of learners no information on expected course completion date was recorded.[16]

[15] This included expansion of response options, moving from a 4-point scale to a 5-point scale, so as to separate out “no use” from “little use”. Some additional questions were also included to cover the role of ICT in teaching and learning, and to ask teachers to assess the overall impact of ICT in learning in class.
[16] We were reliant on providers to obtain this data; despite chasing and visits to providers, the required data were not always received.

5 Sample descriptives, randomisation performance and programme fidelity

In this chapter we present the results of descriptive analysis of the RCT data. We begin by presenting summary statistics on the overall sample. This includes the number of participating centres and learners, the number of learners consenting to participate in various aspects of the study, and the number for whom pre-test and post-test data has been obtained. The chapter then considers how successfully the randomisation worked, with particular emphasis on how well the randomisation achieved two similar-looking groups across the two treatment arms. We also explore whether there were differences in pre-assessment scores by treatment arm, as well as considering time elapsed between learners enrolling and completing the post-assessment. Finally, we consider evidence from teacher responses on how well-differentiated the treatment arms were, as well as whether learners received the mode of learning to which they were assigned.

5.1 Summary statistics

We first present an overview of the RCT sample. We report numbers for both the overall sample, as well as separately for the sample achieved in Phase 1 (the 2013/14 academic year) and Phase 2 (the 2014/15 academic year).

5.1.1 Number of centres participating

Across both phases, a total of 13 centres participated in the study, with four centres participating in both Phase 1 and Phase 2 (Table 5.1). A total of ten centres participated in Phase 1, although in some cases the number of learners recruited from these centres was very small; in three centres, the number of participating learners was less than ten. In Phase 2, a total of seven centres participated.

Table 5.1 Number of participating centres and learners

	Phase 1	Phase 2	Total
Number of centres participating	10	7	13
Number of learners potentially eligible	263	690	953
Number of eligible learners	263	628	891

5.1.2 Number of eligible learners

The achieved sample stood at 953 learners, comprising 263 learners from Phase 1, and 690 learners from Phase 2 (Table 5.1).[17] Not all of these Phase 2 learners were eligible to participate in the study - to be eligible, learners had to be willing and able to attend either the face-to-face learning class or the blended learning class (see Section 3.2). [18] In addition, for a small number of learners eligibility was not known,[19] leaving a total of 891 eligible learners across both phases. While, as noted above, a total of 13 centres participated in the study, there was one large centre that accounted for 42 per cent of all eligible learners.

5.1.3 Number of eligible learners consenting to participate and to data linkage

All eligible learners completed the BQ, with the exception of just one individual in Phase 1 (Table 5.2). Consent to take part in the study and to be randomised was high, at 97 per cent of all eligible learners. Among those individuals who consented to randomisation, 84 per cent (725 learners) gave consent for their information collected within the study to be linked to other information on adult learning (i.e. the ILR). Fewer learners gave consent for their information to be linked to a learner dataset that also includes benefit and employment details (i.e. the linked ILR-WPLS). This possibly reflects a greater sensitivity around such data (the consent question referred not just to employment and benefits but also wages). Lack of consent was a particular issue in Phase 1, where just over half (53 per cent) of those consenting to be randomised gave consent to link to WPLS. It should be noted though that the question requesting consent to link to WPLS was only introduced in November 2013 so the lack of consent does not always imply withheld consent for early participants. The proportion consenting was higher among learners in Phase 2, such that across both phases, 618 learners (72 per cent of those consenting to be randomised) agreed to this data linkage. It is perhaps worth noting that the wording of the consent questions were changed in Phase 2, with the aim of making it easier to understand. This too may have contributed to the higher rate of consent in Phase 2, although it is not possible to tell with certainty.

---

[17] For Phase 1, the original dataset received contained 264 cases. However, one learner appeared to have been randomised twice; all information except the time of randomisation was identical for this learner and so only the first entry was retained. For Phase 2, the original dataset received contained 690 cases. Two cases appeared to be duplicates on all information except eligibility. In both cases, we retained only the eligible records for these learners.

[18] Some learners were also identified as ineligible where they had been randomised for both English and Maths, when they should only have been randomised for one subject, as well as some cases where learners completed the BQ and were randomised multiple times.

[19] There were six cases where eligibility was not known; these cases are considered to be ineligible. In addition, there were two learners who appeared to have been randomised into both arms for the same subject; we consider these cases to be ineligible.

Table 5.2 Number of eligible learners consenting to participate, and number consenting for information to be linked to administrative data sources

	Phase 1	Phase 2	Total
Total number of eligible learners:	263	628	891
Number of eligible learners who:			
...complete the background questionnaire	262	628	890
...consent to randomisation	251	612	863
...consent to take part in follow-up survey*	214	544	758
...consent to match to other information on adult learning	198	527	725
... consent to link to a learner dataset that also includes some benefit and employment details	134	484	618

*the intention initially was to include in the survey all those randomised.

5.1.4 Number of learners by subject

Table 5.3 shows the number of learners participating in the RCT by subject. In both Phases, there were slightly more English than Maths learners. Some learners were studying both English and Maths as part of the RCT, such that some individuals are counted twice in the overall total of 863 learners. Overall, there were 194 learners studying both English and Maths as part of the RCT.[20] Furthermore, some individuals from Phase 2 of the study are counted twice as they re-enrolled for a second course; in which case each course is counted separately. In total this was the case for 53 of the 612 learners in Phase 2 consenting to randomisation. Looking at mode of learning, we see that the numbers of learners are quite evenly balanced across treatment arms for both subjects, particularly in Phase 2.

[20] It is possible that this is an underestimate as this is based on exact matching of learner names.

Table 5.3 Number of randomised learners, by subject and randomisation arm

	Phase 1			Phase 2			Total		
	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total
Number of learners									
Total	132	119	251	306	306	612	438	425	863
English	69	70	139	166	167	333	235	237	472
Maths	63	49	112	140	139	279	203	188	391

5.1.5 Drop out

The rate of drop out is of interest for two reasons. First, as an outcome in its own right; it is interesting to know whether the mode of learning affects how likely individuals are to complete their course. This is particularly relevant with the type and level of learning considered in this evaluation, which is often characterised by high rates of drop out. Second, the degree of drop out affects the proportion of students for which it is possible to administer tests and therefore measure skills. A high rate of drop out reduces the effective sample size at the time of estimating impacts.

Monitoring questionnaires administered to participating centres during the course of the evaluation provide some evidence of drop out. This information was available for those centres participating in Phase 2.

Monitoring records were available for 593 of the 612 eligible learners consenting to randomisation. Among individuals for whom this information was available, 43 per cent were recorded as having withdrawn from the RCT (although this proportion varied across colleges, ranging from 5 per cent to 71 per cent). A variety of factors contributed, including the fact that some learners left their course early, or needed to change day or time of class, and some classes were merged for cost effectiveness reasons. Variation between providers often related to wider issues and policy changes within the organisation, including restructuring, redundancies and staff changes. Among learners, withdrawal from the RCT was generally due to such practical reasons rather than a lack of willingness to be part of the research.

5.1.6 Number of learners for whom we have pre- assessment score

Not all learners who consented to be randomised went on to complete the pre-assessment. In some cases this will reflect drop out from courses, as described above.

Table 5.4 replicates Table 5.3 for those learners for whom pre-assessments are available.

In total, among those learners participating in the RCT, data on completed reading pre-assessments were available for 142 English learners, on writing pre-assessments for 137 English learners, and on maths pre-assessments for 164 learners. For reading and writing, all these individuals come from Phase 2 of the RCT, for maths there are learners from both phases.

Table 5.4 Number of learners completing pre-assessment, by phase, subject and randomisation arm

	Phase 1			Phase 2			Total		
	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total
Number of learners									
English reading	-	-	-	63	79	142	63	79	142
English writing	-	-	-	61	76	137	61	76	137
Maths	24	24	48	54	62	116	78	86	164

Considering this as a proportion of all randomised learners, pre-assessments were available for around three in ten English learners (30 per cent of English learners completed the reading assessment and 29 per cent the writing assessment), and around four in ten maths learners (42 per cent).

5.1.7 Number of learners completing pre- and post-assessments

Even fewer learners went on to complete the post-assessment. Again, to some extent this will also reflect drop out from courses.[21] Table 5.5 shows the number of learners for whom we have both pre and post assessments. In total, among those learners participating in the RCT, 74 completed pre- and post-assessments for reading, 58 did so for writing and 75 for maths. Considering this as a proportion of all randomised learners,

[21] There were some additional learners identified as having completed both pre- and post–assessments, but for whom IRT scores were not available. This applied for four maths learners. Among English learners, 28 did not have IRT scores for the reading assessment and 44 did not have IRT scores for the writing assessment. All were from Phase 2.

both pre- and post-assessments were available for less than one fifth (19 per cent in the case of Maths learners, while 16 per cent of English learners completed the reading assessments and 12 per cent of English learners completed the writing assessments). While this reduces our sample size, drop-out is a common feature of adult learning. Brooks et al. (2008), for example, report that roughly half of adult learners enrolled on literacy courses in England drop out within three months.

Table 5.5 Number of learners completing pre-assessment and post-assessment, by phase, subject and randomisation arm

	Phase 1			Phase 2			Total		
	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total	Face to Face	Blended learning	Total
Number of learners									
English reading	-	-	-	31	43	74	31	43	74
English writing	-	-	-	28	30	58	28	30	58
Maths	12	6	18	27	30	57	39	36	75

Hence, our final estimation sample was made up of 74 English learners with reading assessments, 58 English learners with writing assessments, and 75 Maths learners with maths assessments.

5.1.8 Number of learners for whom we have been able to link to admin data

We attempted to match the RCT data to administrative datasets in order to explore some of the secondary outcomes listed in the protocol. Two ILR datasets were available for the match. The first, for 2013/14, included the following identifying variables:

- Family name
- Given names
- Date of birth
- Address
- postcode
- phone number
- email address.

The second ILR dataset, for 2014/15, includes only:

- Date of birth
- Postcode.

A match was attempted for eligible individuals in the RCT data who consented to their data being linked to the ILR. There were 501 such individuals, some of whom participated in more than one learning aim (and some of whom were randomised twice – only the first randomisation is considered). Table 5.6 shows that, of these learners, 360 could be matched to the ILR. This translates into 256 English learners and 211 maths learners. For many of these individuals, there are multiple records in the ILR, reflecting participation in multiple courses. However, when comparing the start date in the RCT data with that in the ILR, it is often the case that there is no ILR date corresponding to the RCT date. This raises a concern as to whether the learning aim found in the ILR corresponds to the learning aim for which individuals were randomised. To give some sense of this, the table below shows the number of learners who have a match where the dates differ by no more than 30 days. This reduces the sample to a little over 100 for both English and maths learners; roughly 45% of those who could be found in the ILR.

Table 5.6 Number of individuals who could be matched to ILR

	Either subject	
Number of **individuals** consenting to ILR match	501	
Of whom, number found in ILR	360	
	English	**Maths**
Number of **learners** found in ILR	256	211
Of whom, number for which RCT start date within 30 days of start date in ILR	109	102

There are two further points to make. First, we would expect the date of randomisation to come before the ILR learning aim start date since randomisation took place at enrolment. Second, we would expect the courses recorded in the ILR to be English or maths as appropriate. To explore both of these points, the learning aim reference variable in the ILR was linked to a lookup file containing the learning aim description.[22] Where an ILR record could be found for a learner in the RCT (that is, where there was a match), the

[22] The file "LearningDelivery" was used as a lookup table (the file was downloaded from https://data.gov.uk/dataset/learning-aim-reference-service as part of a zip file).

start date from the RCT data, the start date from the ILR and the learning aim title corresponding to that ILR start date could be listed. This revealed two further points:

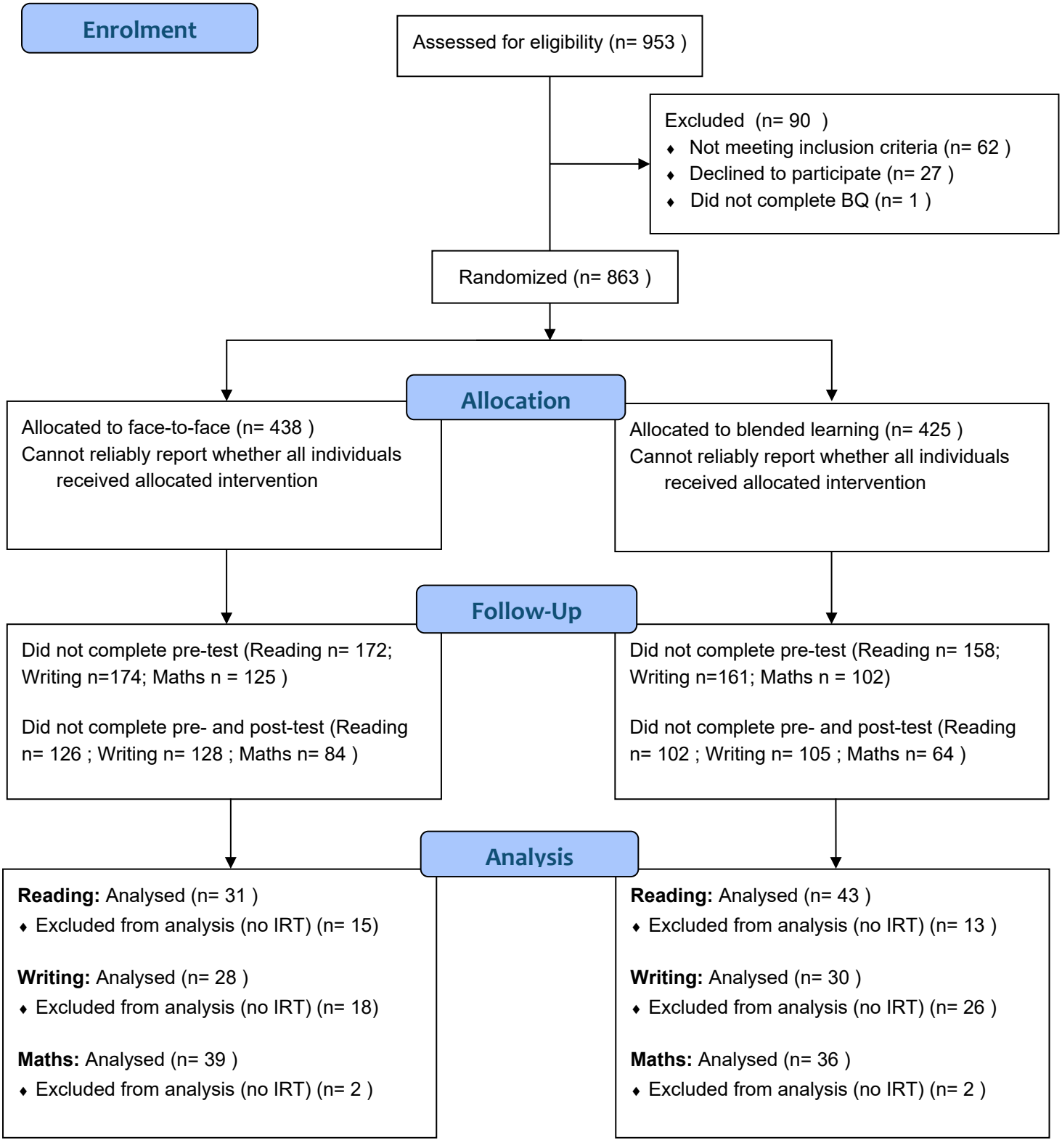
1. There were numerous cases where there was no learning aim recorded in the ILR similar to what was expected for the RCT learner (i.e. either English or maths).
2. Where a match was found, the start date in the ILR was often very different from the randomisation date in the RCT data. Furthermore, among those individuals for whom dates matched were reasonably closely, it was often the case that the randomisation date was later than the start date recorded in the ILR.

These findings reduced the confidence in the quality of the match. Given this concern as well as the small sample size and the fact that the administrative data provided only secondary outcomes, the analysis in this report instead focuses on the primary skills outcome collected through assessment.

5.1.9 CONSORT Diagram

The following CONSORT diagram summarises attrition through the study.

Figure 5.1 CONSORT Diagram



5.2 Examining how well randomisation has worked

5.2.1 Characteristics of sample

We begin by discussing the overall characteristics of individuals who consented to be part of the trial and were successfully randomised. This also allows us to assess whether randomisation has been successful in terms of producing treatment (Blended Learning) and control (Face-to-face) groups that are balanced on observable characteristics. We consider the samples studying English and maths separately.

Participants in the trial had an average age in the mid-thirties (Table 5.7), slightly higher for English (just under 36 years old) than Maths learners (just under 35 years old). There are no statistically significant differences by treatment arm for either subject.

Table 5.7 Age at start of study (years)

	English			Maths		
	Face-to-face	Blended learning	Overall	Face-to-face	Blended learning	Overall
Mean	36.06	35.64	35.85	34.85	34.63	34.74
N	231	233	464	197	188	385
		t= 0.39	p= 0.70		t= 0.19	p= 0.85

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

A large majority of participants in both the English and maths parts of the trial were female (Table 5.8) at over 70%. Among English learners, the blended learning arm had more female participants than the face-to-face arm (a difference of 7 percentage points), although this difference is not statistically significant. The difference between the two arms for Maths learners is much smaller at 4 percentage points.

Table 5.8 Proportion of female participants

	English			Maths		
	Face-to-face	Blended learning	Overall	Face-to-face	Blended learning	Overall
Mean	0.69	0.76	0.72	0.72	0.76	0.74
N	235	237	472	203	188	391
		t= -1.60	p= 0.11		t= -0.70	p= 0.48

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

On average, both maths and English learners left full time education at the age of 19 (Table 5.9). There are only relatively small differences in this characteristic between the blended learning and face-to-face arms. This suggests that average participants had completed more than the UK's level of compulsory education (this may be due to some learners from overseas with qualifications not recognised in the UK).

Table 5.9 Age left full time education

	Face-to-face	Blended learning	Overall
English sample			
Mean	19.41	18.71	19.05
N	181	185	366
		t= 1.28	p= 0.20
Maths sample			
Mean	18.7	18.87	18.78
N	159	151	310
		t= -0.33	p= 0.74

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

The ethnic group distribution of participants was very similar regardless of subject studied across the two arms of the trial in each case (Table 5.10). The largest single group of participants were White, at just under 40%, followed by individuals who were Black, making up just over 30% of participants. Comparing this to census data for 2011 we find that individuals of non-white ethnicity are heavily over-represented relative to their share

of the population,[23] which is approximately 14% of the English and Welsh population compared with about 60% in this trial. This is more representative of London, which is perhaps unsurprising given the large proportion of the trial's participants attending one large centre in London (as mentioned in Section 5.1.2). There is no difference in ethnic group distribution by treatment arm for either subject.

[23] http://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/2011censuskeystatisticsforenglandandwales/2012-12-11#ethnic-group

Table 5.10 Ethnicity, column percentages

	Face-to-face	Blended learning	Overall
English sample			
White	40	37.1	38.6
Mixed	3.4	4.2	3.8
Asian	13.6	16.9	15.3
Black	30.6	30.8	30.7
Other	11.1	8	9.5
Prefer not to say	1.3	3	2.1
Total	100	100	100
N	235	237	472
	Pearson chi2(5) = 3.9963 Pr = 0.550		
Maths sample			
White	37.4	37.8	37.6
Mixed	5.9	3.7	4.9
Asian	14.8	16	15.3
Black	32	30.9	31.5
Other	6.9	9.6	8.2
Prefer not to say	3	2.1	2.6
Total	100	100	100
N	203	188	391
	Pearson chi2(5) = 2.2120 Pr = 0.819		

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

Turning next to participants' partnership statuses (Table 5.11), somewhat over half of participants were single, with the vast majority of the remainder being in partnerships of various kinds (including marriages, civil partnerships and living with a partner). A small group preferred not to give an answer. Nevertheless, there is no evidence that partnership status varies by treatment arm; this finding holds when partnership groups are not aggregated into these broader groups.

Table 5.11 Partnership status, column percentages

	Face-to-face	Blended learning	Overall
English sample			
Single	58.3	56.5	57.4
Partnered	37	36.7	36.9
Prefer not to say	4.7	6.8	5.7
Total	100	100	100
N	235	237	472
			Pearson chi2(2) = 0.9507 Pr = 0.622
Maths sample			
Single	56.7	54.3	55.5
Partnered	37.9	42	39.9
Prefer not to say	5.4	3.7	4.6
Total	100	100	100
N	203	188	391
			Pearson chi2(2) = 1.1195 Pr = 0.571

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

Again, there is no evidence of difference in the economic activity of participants depending on the treatment arm to which they had been assigned (Table 5.12). The largest single group were the economically inactive, followed by individuals in part time employment. This is perhaps unsurprising given that many participants reported that they wanted to take their course in order to help them get a job or to help them get a better job.

Table 5.12 Economic activity, column percentages

	Face-to-face	Blended learning	Overall
English sample			
Full time Work	10.6	12.2	11.4
Part time Work	27.7	20.7	24.2
Unemployed	9.8	10.5	10.2
Economically Inactive	31.9	39.7	35.8
Other	12.8	12.2	12.5
Prefer not to say	7.2	4.6	5.9
Total	100	100	100
N	235	237	472
			Pearson chi2(5) = 6.0556 Pr = 0.301
Maths sample			
Full time Work	10.3	14.4	12.3
Part time Work	26.6	25	25.8
Unemployed	8.9	8	8.4
Economically Inactive	37.9	31.9	35
Other	10.8	14.4	12.5
Prefer not to say	5.4	6.4	5.9
Total	100	100	100
N	203	188	391
			Pearson chi2(5) = 3.6009 Pr = 0.608

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals who are randomised in both English and maths parts of the trial appear in both sections of the table.

Overall, in this section we have reported some of the key demographic characteristics of participants in the trial. We have also found that the initial randomisation appears to have been successful in producing well-balanced treatment and control groups in terms of observable characteristics. We move on to consider how the characteristics of the sample have changed as a result of attrition, first before pre-tests have been completed and then before post-tests have been completed, including whether this attrition has resulted in differences between the treatment and control arms.

5.2.2 Characteristics of learners for whom we have pre-tests

Not all individuals eligible for the trial and successfully randomised went on to complete a pre-test. As such, the potential sample for the impact analysis of this trial has been reduced. In addition, some individuals in the English sample completed a reading, but not a writing assessment or vice-versa; the differences are small at this point, but become larger when we consider individuals for whom we also have post-tests. From this point onwards, we discuss reading, writing and maths samples separately. At this point, the reading sample is reduced from 472 randomised English learners to 142 with valid reading pre-tests. Also, the writing sample is reduced from 472 randomised English learners to 137 with valid writing pre-tests and the maths sample is reduced from 391 randomised maths learners to 163 with valid maths pre-tests.

The implications of reduced sample size are that statistical power is reduced (that is, the RCT becomes less able to detect an effect of a given size) and that those remaining in the sample may not be representative of the randomised population. This affects how representative the resulting estimates can be felt to be; that is, their 'external validity'. In addition, it is possible that the failure to complete the pre-test is not only non-random but also varies by treatment group. In this case, the internal validity of the trial is reduced. In other words, should characteristics of learners vary across arms of the trial, we can no longer be confident in the ability of the trial to provide robust impact estimates.

We repeat the balancing analysis for this restricted sample of learners who completed the relevant pre-tests; the full results are presented in Appendix B.1. Summarising, we find no evidence of imbalance between treatment and control groups. This does not confirm definitively that there are no systematic differences between the arms – there could, after all be unobserved differences – but it is a reassuring finding and gives more confidence in the ability of the sample to provide unbiased impact estimates.

5.2.3 Characteristics of learners for whom we have pre-tests and post-tests

Not all individuals who completed a pre-test went on to complete a post-test. This further reduces the sample size which is useable for the impact estimates in this report. The reading sample is reduced from 472 randomised English learners to 74 with valid reading pre- and post-tests. The writing sample is reduced from 472 randomised English learners to 58 with valid writing pre- and post-tests and the maths sample is reduced from 391 randomised maths learners to 75 with valid maths pre- and post-tests.

As with attrition before pre-testing, it is possible that the failure to complete the post-test is non-random and, hence, the composition of the sample has changed in ways that reduce the internal validity of the trial. The balancing analysis is repeated for this

restricted sample of learners completing pre- and post-tests, with full results presented in Appendix B.2.

Again, the overall finding is that we still have treatment and control groups which are not statistically significantly unbalanced on observable characteristics. As this is the sample used to calculate impact estimates, this gives some level of reassurance regarding the internal validity of our results.

5.2.4 Mean levels of pre-tests and distributions by treatment arm

A particularly important aspect of balance between the two arms of the trial is performance in the pre-test. In this section, we explore how pre-test performance is balanced between treatment arms and how this changes when we restrict attention to only those who also have valid post-tests. We consider both the mean levels of performance, but also the overall distribution. As with the other balancing tests, we assess the reading, writing and maths samples separately.

Individuals' reading and maths scores were estimated using a Rasch/one parameter Item Response Theory (IRT) model. An important point to note is that since the aim of this modelling is to put all individuals' scores on a common scale it is necessary to provide a distribution for this new scale. In this case, the scale is centred on zero with a standard deviation of one, so that all differences in scores may be interpreted as differences in standard deviations of overall performance by learners. Boyle and Horrocks (2015) provide details of the approach and IRT more broadly. They describe how scores for learners at different levels were put on a single scale. We note that this approach does not guarantee the resulting overall sample mean and standard deviation will equal zero and one, respectively. For writing scores a different – hybrid – approach is used: IRT for spelling, punctuation and grammar and a simple continuous score for extended writing.

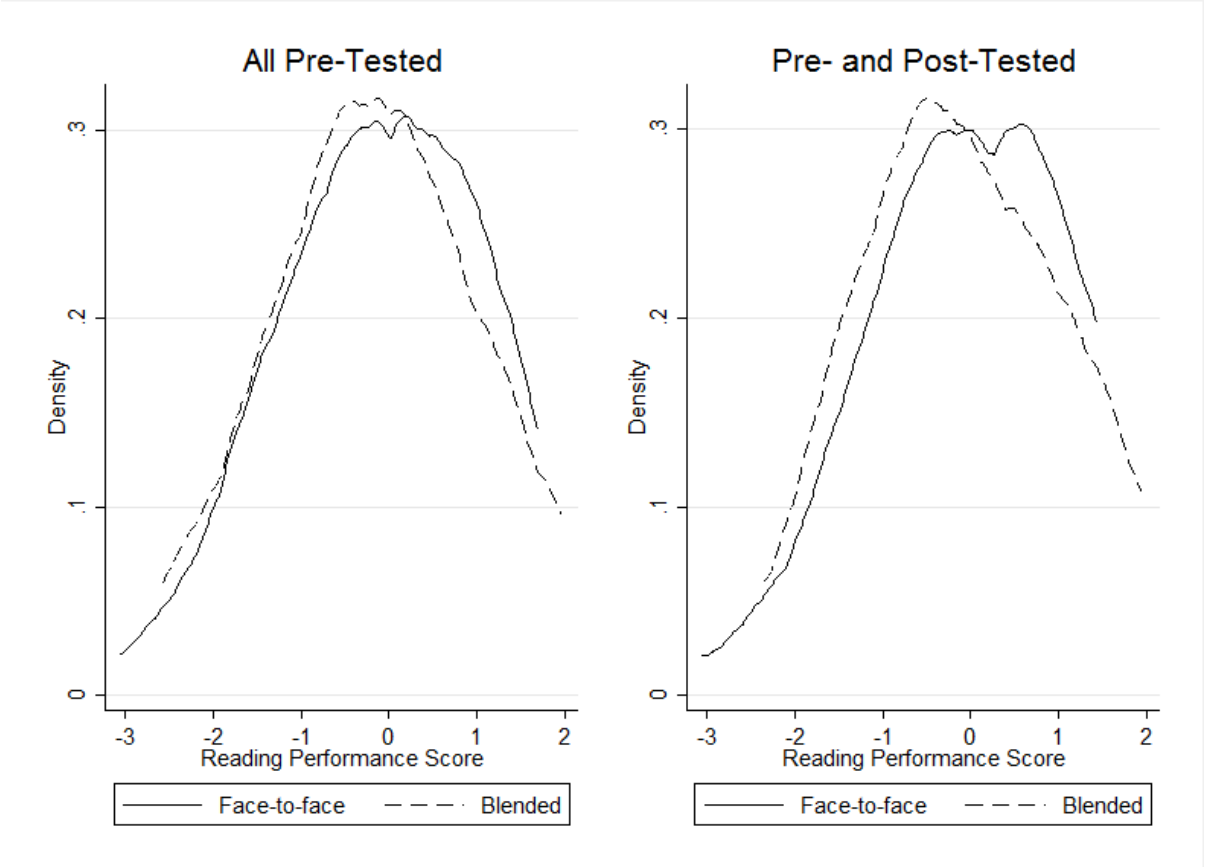
The average performance in the reading test (Table 5.13), among all those who took it, was -0.13, with no significant difference by whether participants were assigned to the face-to-face or the blended learning arms. Among those who also took the post-test, the average performance is -0.08, a very slightly higher performing sample. In both cases, individuals in the face-to-face arm have slightly better performance, but there is no evidence of a statistically significant difference by treatment arm. The same message is evident in plots of the distribution of reading pre-test performance (Figure 5.2), with the slightly higher performance of the face-to-face participants evident across the distribution and slightly more evident among those who completed both pre- and post-tests; the difference is, nevertheless, not statistically significant.

Table 5.13 Mean reading pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample

	Face-to-face	Blended learning	Overall
Pre-tested sample			
Mean	-0.12	-0.15	-0.13
N	63	79	142
		t= 0.13	p= 0.90
Pre-tested & post-tested sample			
Mean	-0.04	-0.11	-0.08
N	31	43	74
		t= 0.29	p= 0.78

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant i) pre-test and ii) pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups.

Figure 5.2 Distribution of reading pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample



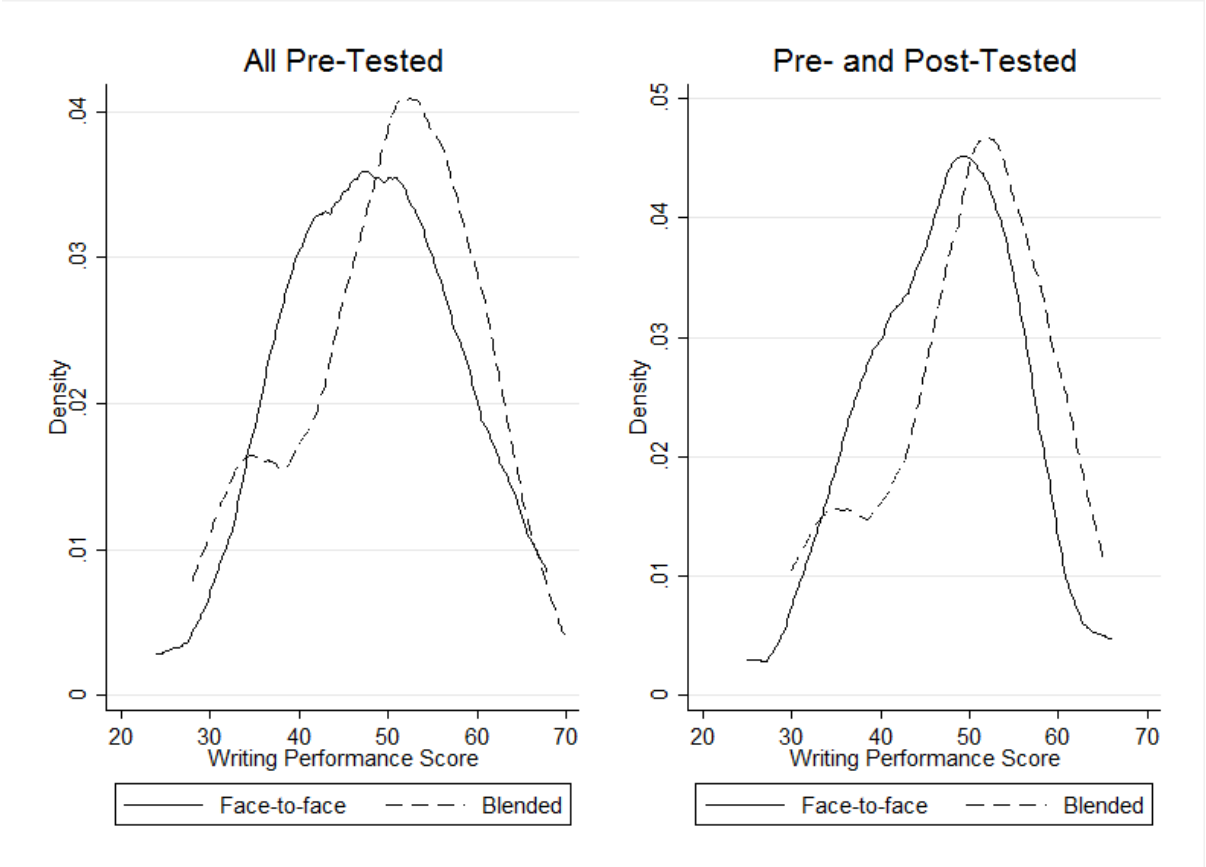
The average performance in the writing test (Table 5.14), among all those who took it, was 49.0, with no significant difference by whether participants were assigned to the face-to-face or the blended learning arms. Among those who also took the post-test, the average performance is 48.38. In both cases, individuals in the blended learning arm have slightly better performance, but there is no evidence of a statistically significant difference by treatment arm. Plots of the distribution of writing pre-test performance (Figure 5.3) demonstrate more difference in the distributions than was evident in the case of reading, although this is no more pronounced among those that took both the pre- and post-tests rather than just the pre-test.

Table 5.14 Mean writing pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample

	Face-to-face	Blended learning	Overall
Pre-tested sample			
Mean	48.21	49.63	49
N	61	76	137
		t= -0.84	p= 0.40
Pre-tested & post-tested sample			
Mean	46.89	49.77	48.38
N	28	30	58
		t= -1.23	p= 0.22

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant i) pre-test and ii) pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups.

Figure 5.3 Distribution of writing pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample



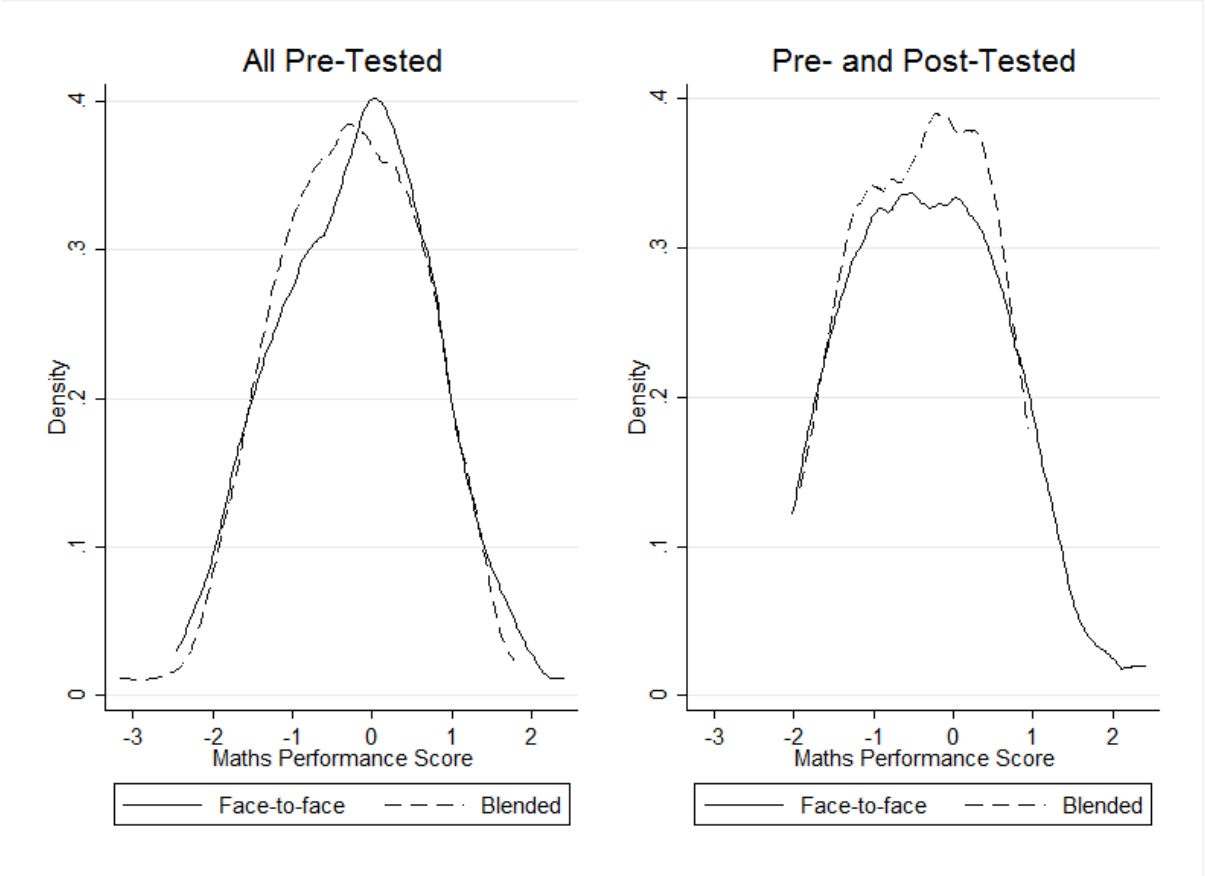
The average performance in the maths test (Table 5.15), among all those who took it, was -0.25, with no significant difference by whether participants were assigned to the face-to-face or the blended learning arms. Among those who also took the post-test, the average performance is -0.36; as with the writing test this is slightly worse performance than in the broader sample. In both cases, individuals in the face-to-face arm have slightly better performance, but there is no evidence of a statistically significant difference by treatment arm. Plots of the distribution of maths pre-test performance (Figure 5.4) demonstrate that these are fairly similar between the two arms, although there is some change between the full pre-tested sample and those who took the pre- and post-tests.

Table 5.15 Mean maths pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample

	Face-to-face	Blended learning	Overall
Pre-tested sample			
Mean	-0.2	-0.29	-0.25
N	78	86	164
		t= 0.59	p= 0.55
Pre-tested & post-tested sample			
Mean	-0.32	-0.41	-0.36
N	39	36	75
		t= 0.39	p= 0.70

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant i) pre-test and ii) pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups.

Figure 5.4 Distribution of maths pre-test by treatment arm for i) pre-tested sample and ii) pre-tested and post-tested sample



5.2.5 Time from randomisation until post-test

Finally, for this section, we consider whether there are differences between the treatment arms in the length of time individuals took to complete the course and how long after starting the course they took a post-test. As discussed in the interim report, there was some evidence of difference in the length of time to maths pre-test between the face-to-face and blended learning arms (Anders, Dorsett and Stokes, 2015).

Figure 5.6 shows the proportion of English learners who have not completed a reading post-test as the time since randomisation lengthens. Figure 5.6 repeats this for maths. In both cases, the differences between treatment arms look rather slight and indeed statistical tests give no reason to believe there are any systematic differences.

Figure 5.5 Time from randomisation to reading post-test by treatment arm

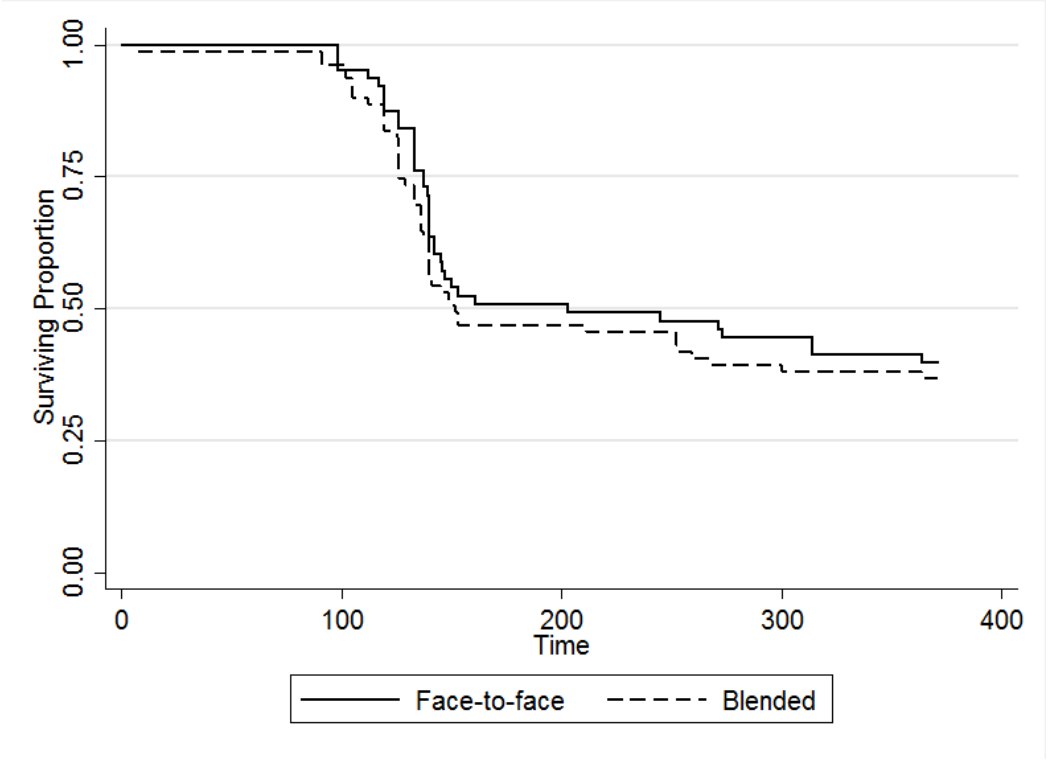
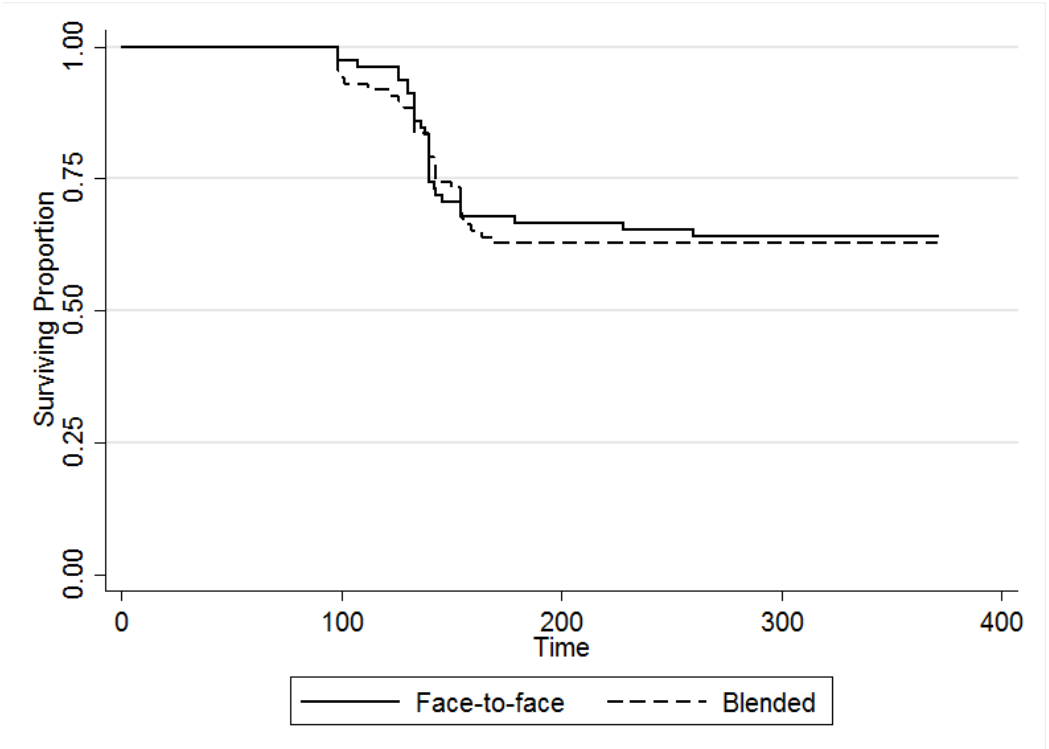


Figure 5.6 Time from randomisation to maths post-test by treatment arm



5.3 Programme fidelity

5.3.1 Teacher assessments of the extent of ICT use

As discussed in Section 2.1, a key factor for the success of the RCT is the extent to which the two modes of learning can be clearly distinguished from one another. In this section we explore this issue using responses received from the teacher questionnaire. The Phase 2 teacher questionnaire[24] asked teachers to assess the extent to which ICT was used in their class along 13 different dimensions. The first four of these dimensions related to the extent to which ICT was a fundamental part of teaching and learning, as well as learners' active and passive use of ICT in their learning; these were considered to be the four key measures.[25] The remaining questions focus on more specific examples of the different uses of ICT in teaching and learning (Murphy, Grant and Smith, 2014). The questionnaire also collects additional information on the class and teacher, including class subject, level, and start and finish dates. It also asks specifically whether the class is a blended learning or face-to-face class. Teachers were asked to complete the questionnaire at three points in time, at the start of learning for the class, at the middle, and at the end.

In total, 124 responses to the Phase 2 teacher questionnaire were submitted. At least one response was received from each of the 7 colleges participating in Phase 2 of the RCT, although over half of responses came from one college. In all, 38 responses related to the start of the course, 43 responses to the middle of the course, and 40 to the end of the course (the remaining 3 responses indicated other learning points). The responses were fairly evenly distributed by subject (with a total of 63 responses for maths and 61 responses for English), and by treatment arm (62 relating to blended learning classes and 58 to face-to-face classes, with 4 classes where this information was missing).

Earlier analysis of responses to the teacher questionnaire (based on the smaller number of responses received at the time) confirmed the distinctiveness of the two treatment arms (Murphy, Grant and Smith, 2014; Anders, Dorsett and Stokes, 2015). This continues to hold when extending this analysis to incorporate the full set of responses received.

Table 5.16 reports the mean scores on each of the four key dimensions assessing the extent of ICT use. Each response is scored on a scale from 1 to 5, with a higher score indicating higher use of ICT. For all four dimensions, the average score is higher for blended learning than face-to-face classes. Table 5.16 also reports the results of constructing two summary measures, the first calculated as an average of the responses

[24] As noted in Section 4.6, the teacher questionnaire was revised between Phase 1 and Phase 2 of the RCT. Only responses from Phase 2 have been used in this analysis.
[25] These were determined through Delphi group work with an expert group.

on the four core dimensions, and the second as an average across all 13 dimensions. Both of these measures also suggest a clear difference in the extent of ICT use between treatment arms.

The relatively small sample sizes limit the feasibility of analysis by subgroup. However, differences in mean scores do also appear to be apparent by treatment arm for each subject (Table 5.16), and for each learning point.

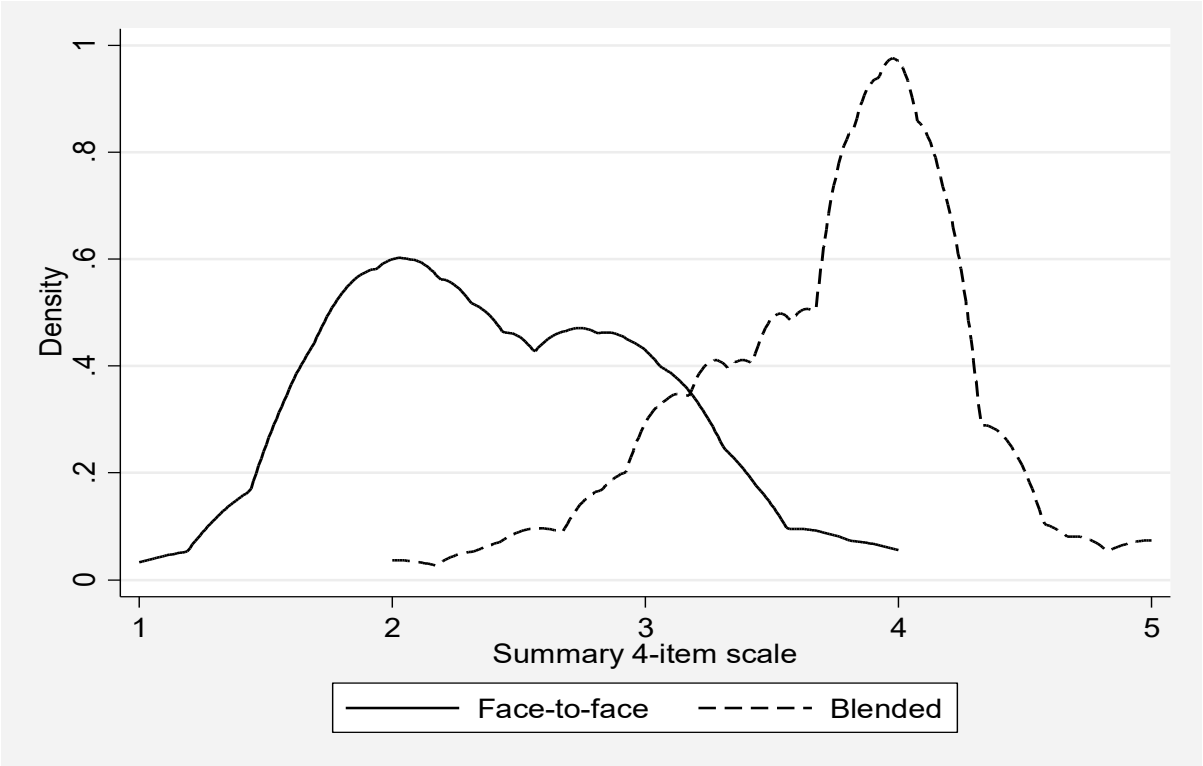
It should be noted that these figures focus on average scores. There is of course some variation, such that some blended learning classes have lower ICT usage than face-to-face classes, according to these measures (Figure 5.7). For example, on the summary 4-item measure, scores for face-to-face classes ranged from 1 to 4, while for blended learning this ranged from 2 to 5.

Table 5.16 Mean scores on teacher questionnaire by treatment arm	Face-to-face	Blended learning	Overall
ICT use by learners as a fundamental aspect of their learning activities			
Mean	2.53	3.92	3.25
N	58	62	120
		t=10.54	p=0.00
ICT use by teachers and other learning support staff as a fundamental aspect of their work with learners			
Mean	2.43	3.79	3.13
N	58	62	120
		t=10.39	p=0.00
Extent to which learners are 'consuming' ICT			
Mean	2.50	3.66	3.10
N	58	62	120
		t=10.14	p=0.00
Extent to which learners are 'actively doing' in relation to ICT use			
Mean	2.24	3.56	2.93
N	58	62	120
		t=8.80	p=0.00
Average score: 4 items			
Mean	2.43	3.73	3.10
N	58	62	120
		t=12.00	p=0.00
Average score: 13 items			
Mean	2.06	3.25	2.67
N	58	62	120
		t=11.18	p=0.00
Average score: 4 items (English)			
Mean	2.53	3.73	3.13
N	30	30	60

		t=7.94	p=0.00
Average score: 4 items (Maths)			
Mean	2.32	3.73	3.08
N	28	32	60
		t=9.03	p=0.00
Average score: 13 items (English)			
Mean	2.18	3.22	2.70
N	30	30	60
		t=7.38	p=0.00
Average score: 13 items (Maths)			
Mean	1.93	3.27	2.65
N	28	32	60
		t=8.49	p=0.00

Notes. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning classes.

Figure 5.7 Distribution of summary 4-item measure, by treatment arm



5.3.2 Evidence on whether teachers adhered to assigned mode

The previous subsection compared mode of learning across face-to-face and blended classes, as reported by teachers. There is a question around whether learners received the mode of learning to which they were assigned. Using the monitoring data, for a subset of learners where information is available, it is possible to compare individuals' randomisation outcomes - that is, the mode of learning they were assigned to receive - against the mode of learning they actually experienced. Of the 299 eligible and randomised learners for whom this information was available, 91 per cent were receiving the form of learning to which they had been assigned. Table 5.17 shows that a small number of learners appear to be in the "wrong" arm. Such non-compliance is a routine feature of RCTs. We note that the number of non-compliers is fairly low, suggesting that this is unlikely to materially alter the results.

Table 5.17 Distribution of learners by assigned and actual treatment arm

		Actual		
		Face-to-face	Blended learning	Total
Assigned	Face-to-face	140	13	153
	Blended learning	15	131	146
	Total	155	144	299

6 Estimated impacts

In this section, we estimate the impact of individuals receiving blended learning rather than face-to-face tuition on their reading, writing and maths skills. This is provided as a difference in IRT scores (in the case of reading and maths) or total scores (in the case of writing). Given the smaller than initially targeted sample, we also provide estimates of the minimum detectable effect size from the sample that was available. This provides context to the results.

The results are estimated using linear regression models. We estimate two models for each outcome variable. The first includes as a covariate only the college in which individuals complete their course.[26] In order to increase the precision of our estimates (reducing standard errors and increasing power), we estimate a second set of models that include the individual's pre-test performance along with a number of covariates from the background questionnaire (gender, age, ethnic group, economic activity, self-reported IT confidence, whether English is an Additional Language, previous highest level of English qualifications, and previous highest level of Maths qualifications).

We do not conduct any sub-group analysis; given the small sample sizes available for the overall sample it would not yield any insightful results. However, the covariates included in the regression model are based on those identified as potentially interesting for sub-group analysis in the evaluation protocol.

In order to provide our results in a form that is comparable with other studies, we convert the estimated effects into effect sizes using the procedure suggested by Hedges (1981) and refined by Hedges and Olkin (1985). This places the difference in units of the pooled standard deviation of the sample (further details are given in Appendix C). An effect size of 1 therefore corresponds to an impact of 1 standard deviation. This would typically be viewed as a large effect, sufficient to correspond to a shift from the 31st percentile of the outcome distribution to the 69th percentile. We also use the relevant features of the achieved sample (notably sample size and predictive power of any covariates used) to estimate the minimum detectable effect size with 0.80 power for significance tests at the 0.05 level for each analysis. This may be interpreted as the smallest true effect size that we would have an 80% chance of finding to be statistically significant at the 5% level given the available sample.

[26] These are included to account for centre-specific effects and reflect the fact that randomisation took place within centres.

6.1 English – Reading

6.1.1 Without covariates

We first estimate the impact of blended learning (treatment) relative to face-to-face (control) on reading scores (Table 6.1). Those who attended blended learning courses had reading post-test scores of 0.22, relative to post-test scores of 0.05 for those who attended face-to-face courses. The difference between their performance is 0.16 (note there is a slight difference due to rounding), but this is far from statistically significant. This converts to an effect size of 0.15, which would be a meaningful difference were it significant. Using the realised features of the sample, we estimate that the minimum detectable effect size (MDES) for this specification would be 0.48; this highlights that the non-significance of our finding is unsurprising given the lack of power.

Table 6.1 Estimated impact on reading scores (no covariates)

	Estimate	Standard Error
Face-to-face	0.05	0.15
Blended learning	0.22	0.17
Difference	0.16	0.23
Effect size	0.15 [0.65]	
MDES	*0.48*	

Notes. Results from linear regression of post-test score on treatment indicator and college dummy variables. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

6.1.2 With covariates

Adjusting for the covariates (Table 6.2), we find a larger effect size of 0.22, but this is still far from statistically significant. Based on the predictive power of the covariates and, again, other relevant details of the realised sample, we estimate that the minimum effect size with the achieved sample size to be 0.29.

Table 6.2 Estimated impact on reading scores (with covariates)

	Estimate	Standard Error
Face-to-face	0.01	0.12
Blended learning	0.25	0.17
Difference	0.24	0.24
Effect size	0.22 [0.96]	
MDES	*0.29*	

Notes. Reporting results from linear regression model of post-test score on: treatment indicator, pre-test score college dummy variables, gender, age, ethnic group, economic activity, self-reported IT confidence, whether English is an Additional Language, previous highest level of English qualifications, and previous highest level of Maths qualifications. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

6.2 English – Writing

6.2.1 Without covariates

Turning to writing scores (Table 6.3), those who attended blended learning courses had writing post-test scores of 52.28, relative to post-test scores of 51.41 for those who attended face-to-face courses. The difference between their performance is 0.87 points, but this is once again far from statistically significant. This converts to an effect size of 0.09, a smaller, but still potentially meaningful difference were it significant. The estimated MDES from the achieved sample is even larger than was the case for reading performance at 0.53.

Table 6.3 Estimated impact on writing scores (no covariates)

	Estimate	Standard Error
Face-to-face	51.41	1.71
Blended learning	52.28	1.43
Difference	0.87	2.24
Effect size	0.09 [0.36]	
MDES	0.53	

Notes. Reporting results from linear regression model of post-test score on treatment indicator and college dummy variables. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

6.2.2 With covariates

Adjusting for the covariates (Table 6.4), we find a larger effect size of 0.22, but this is still far from statistically significant. Based on the predictive power of the covariates and other relevant details, we estimate the minimum effect size that would have been detectable with the achieved sample size to be 0.31. As with reading, it is clear that there is not sufficient power to detect effect sizes of the magnitude observed with the achieved sample. These results for writing therefore resemble those for reading.

Table 6.4 Estimated impact on writing scores (with covariates)

	Estimate	Standard Error
Face-to-face	50.82	2.01
Blended learning	52.84	1.5
Difference	2.02	3
Effect size	0.22 [0.84]	
MDES	0.31	

Notes. Reporting results from linear regression model of post-test score on treatment indicator, pre-test score, college dummy variables, gender, age, ethnic group, economic activity, self-reported IT confidence, whether English is an Additional Language, previous highest level of English qualifications, and previous highest level of Maths qualifications. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

6.3 Maths

6.3.1 Without covariates

Finally, we consider the case of maths (Table 6.5). Those who attended blended learning courses had reading post-test scores of 0.08, relative to post-test scores of 0.05 for those who attended face-to-face courses. The difference between their performance is 0.04 (note that there is a slight difference due to rounding), but this is far from statistically significant. This converts to an effect size of 0.04, which is unlikely to be substantively meaningful, even if it were statistically significant.

Table 6.5 Estimated impact on maths scores (no covariates)

	Estimate	Standard Error
Face-to-face	0.05	0.12
Blended learning	0.08	0.14
Difference	0.04	0.19
Effect size	0.04 [0.18]	
MDES	0.41	

Notes. Reporting results from linear regression model of post-test score on treatment indicator and college dummy variables. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

6.3.2 With covariates

Adjusting for the covariates (Table 6.6), we find a larger effect size of 0.10, but this is still far from statistically significant. Based on the predictive power of the covariates and other relevant details, we estimate that the minimum effect size that would have been detectable with the achieved sample size would have been 0.25. As with all specifications considered, the estimated effect size falls well short of the minimum detectable effect size of the achieved sample, making it impossible to make statements about the results in which we can be confident.

Table 6.6 Estimated impact on maths scores (with covariates)

	Estimate	Standard Error
Face-to-face	0.02	0.10
Blended learning	0.11	0.12
Difference	0.09	0.17
Effect size	0.10 [0.44]	
MDES	0.25	

Notes. Reporting results from linear regression model of post-test score on: treatment indicator, pre-test score, college dummy variables, gender, age, ethnic group, economic activity, self-reported IT confidence, whether English is an Additional Language, previous highest level of English qualifications, and previous highest level of Maths qualifications. Effect size is Hedges g*, with t-statistic in brackets. Estimated minimum detectable effect size is reported for comparison with estimated effect size.

7 Conclusion

The primary aim of this RCT was to test whether blended learning and face-to-face learning differ in their effectiveness at increasing skills. This is an important question and having an insight into the consequences of shifting away from a traditional pedagogy and more towards one where technology plays a central role can help inform whether this is a positive development. A priori, this is uncertain. One perspective might be that ICT can complement the teacher's role, enhancing productivity and broadening the expertise accessible by the learner. A different view might be that teachers are better able to provide a learning environment that is sensitive to the needs of students and so conducive to learning.

The initial challenge in carrying out the RCT was to recruit enough colleges to allow a sufficient number of learners to be randomised. The learners themselves tended to be happy to participate. In the event, about 900 eligible learners were randomised, considerably below the target of 1,500, but still a sufficient number to provide a reasonable level of statistical power. However, partly due to a high level of drop-out, the proportion of learners for whom skills were assessed at both course start and course end was low. This is relevant since estimation relied on observing skills at both points. In fact, reading skills were assessed at both points for only 74 learners, writing skills for 58 learners and maths skills for 75 learners. This represents substantial sample loss and greatly reduces the ability of the RCT to detect effects.

While this is obviously unfortunate, in other regards the RCT seems to have worked well. Randomisation appears to have been successful in achieving two groups of individuals who look similar in terms of background characteristics and skills as assessed at the start of the course. This similarity is evident whether considering all those randomised or only those for whom both pre- and post-tests are available. Equally important, information collected from teachers confirms that there were meaningful differences between the treatment arms in how learning was delivered. Levels of ICT use among the blended learning classes were higher than among the face-to-face classes. There is also variation; those face-to-face classes with the highest ICT element appear, on the basis of the teacher questionnaire, to offer a more technology-enhanced pedagogy than those blended classes with the lowest ICT element. Such variation is unsurprising, particularly given that nearly all face-to-face classes rely on ICT to some extent. Some non-compliance in the form of learners assigned to one treatment arm receiving the treatment associated with the other treatment arm is also evident. Again, this is common in RCTs and, as far as we can tell, is not widespread in this case.

In view of this, it is worth considering the estimated impact estimates. For English learners, the estimated effect size of blended learning compared to face-to-face learning is 0.22. This holds for both reading and writing. However, it is not statistically significant. While the design of the experiment was sufficient to detect an effect of this size, the

achieved sample is too small. To provide some context, the results show also the minimum detectable effect size with the achieved estimation sample. For reading this is 0.29, for writing it is 0.31. This means that to reliably detect a significant effect, it would have to be of that order. The corollary to this is that we cannot rule out that blended learning affects outcomes in a meaningful way – an effect size of about 0.3 is not small, particularly when considering relative effectiveness – but we have too few observations to register it. For maths learners, the estimated effect size is 0.10 as compared to a minimum detectable effect size of 0.25.

Hence, for both subjects, the consequence of the RCT being underpowered is that the results are rather inconclusive. Furthermore, we should be cautious about regarding these non-significant results as even indicative. They are subject to considerable random variation and there is no guarantee that bigger samples would give results in any way similar to those found with the reduced samples available here.

This is clearly a disappointing outcome since it does not advance our understanding of the effect of mode of learning. Realistically, it is a reflection of the difficulty of the evaluation task in this case.

Phase 1 of the RCT demonstrated some of the difficulties in implementation. A number of changes were made in Phase 2 in order to address these. This included concentrating on a smaller number of providers, but who could offer a sizeable number of learners; and, devolving responsibility for some aspects of project management to the providers, and offering additional funding. However, it is clear that a number of broader challenges remained in Phase 2. Fundamentally, it proved difficult to achieve the sustained learner participation required to allow skills measurement at both the start and end of the course.

Nevertheless, there are a number of lessons to be learned from this study:

- The RCT was resource intensive for providers. They lacked the capacity to take on the additional work that the RCT presented, especially when facing many competing demands. Providers required substantial external support, and would have required considerable financial incentives to enable their existing staff to have the necessary time to take part in the research. However, even when offered, not all providers took up the offer of additional help.
- Designing the research in order to minimise the additional burden on providers is key. This includes careful consideration of requirements for data collection, which in this case had proved burdensome, The assessments were also lengthy, taking up substantial and valuable class time; it is important to strike a balance between assessments that are robust and fit for purpose against the practicalities of administration.

- Many providers were not ready to deliver blended learning to the extent that the RCT had envisaged – not all providers, for example, had the necessary infrastructure and technology in place. The timeline for the evaluation was very ambitious and putting in place the necessary arrangements to deliver the trial from the start of the 2013/14 academic year proved too great a challenge for many providers.
- As far as is feasible, the RCT needs to fit alongside the practicalities of the day-to-day running of provision. In this study for example, there was a tension between the need to randomise learners to the different modes of learning and offering learners flexibility and choice.
- To be a success an RCT requires commitment from all parties involved. It is encouraging that senior leaders were enthusiastic about participating in the research. However, in some cases teachers and learners needed reassurance about the motivation behind the study as well as other forms of support, such as adequate time for training in blended learning for teachers.

Many of the above points are applicable to the design and delivery of RCTs both within and beyond the FE sector: ensuring burdens on participants created by the need for data collection are kept to a minimum, using appropriate but practical means of assessment, and a need for commitment and engagement among all participants. Finally, it is worth noting that qualitative interviews with those taking part in the research indicated that learners and providers considered that learning approaches that included both face-to-face and use of information learning technologies provided an effective learning experience. Regardless of whether learning was face-to-face or blended, learners valued having proximity to a teacher.

While there are undoubtedly lessons to be learned, we should not draw the general conclusion that RCTs have no role in providing evidence in the area of skills. There were several aspects to this RCT that made it particularly challenging. Furthermore, these were apparent from the start: the randomisation itself was not costless for colleges to accommodate; the logistics of administering large numbers of assessments was very demanding; and, the problem of high drop-out rates among adult learners is well-known. Another RCT may well have characteristics that raise many fewer difficulties.

References

Anders, J., Dorsett, R. and Stokes, L. (2015) The relative effectiveness of blended vs. face-to-face delivery of adult English and maths training: Interim report, March 2015.

Bernard, R., Abrami, P., Lou, Y., Borokhovski, E., Wade, A., Wozney, L., Wallet, P., Fiset, M. and Huang, B. (2004) "How Does Distance Education Compare with Classroom Instruction? A Meta-Analysis of the Empirical Literature," Review of Educational Research 74 (2004): 379–80.

Brooks, G., Burton, M., Cole, P., Miles, J., Torgerson, C. and Torgerson, D. (2008) "Randomised controlled trial of incentives to improve attendance at adult literacy classes" Oxford Review of Education 34(5).

Means, B., Toyama, Y., Murphy, R., Bakia, M. and Jones, K. (2010) *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies* U.S. Department of Education Office of Planning, Evaluation, and Policy Development, Washington, D.C., 2010.

Hattie, J. (2009) *Visible learning* London: Routledge

Appendices

A. Research protocol

Background

Significance

The Coalition Government announced in *Skills for Sustainable Growth*[27] (November 2010) that it would continue to fund literacy and numeracy courses for adults who lack basic literacy and numeracy skills, but that, in order to maximise the economic and personal returns from this investment, the Department for Business, Innovation and Skills (BIS) would review the way this provision is delivered and make it more effective. This review was undertaken in 2011 and the outcomes were published in *New Challenges, New Chances*[28] (December 2011). It included a review of the available research and evaluation evidence, also published in December 2011. [29] This identified a "lack of good evidence on information and communications technology (ICT), and on the role and impact of ICT in blended learning provision".

The purpose of this impact analysis is to assess the relative effectiveness of face-to-face learning compared to blended learning.

Intervention

The analysis considers the relative effectiveness of two modes of learning. In order to be able to deliver informative results, it relies on there being a meaningful distinction between face-to-face learning and blended learning. The two models of learning are distinguished by the extent and nature of their use of technology.

- With regard to the **extent** of use, we interpret
    - face-to-face learning as using technology for less than 5 per cent of guided learning hours
    - blended leaning using technology for at least one third of guided learning hours.
- With regard to the **nature** of the use of technology:
    - Face-to-face learning may include an element of technology to practise or consolidate skills through self-study but this must not form a significant part

[27] http://www.bis.gov.uk/policies/further-education-skills/skills-for-sustainable-growth
[28] http://www.bis.gov.uk/newchallenges
[29] http://www.bis.gov.uk/assets/biscore/further-education-skills/docs/r/11-1418-review-research-on-improving-adult-skills

of the course. Tutors may use the web for ideas, activities and resources to use in their face-to-face teaching and the class may use computers for one-off sessions but this should be teacher-led.

- Blended learning combines multiple delivery media that are designed to complement each other and promote learning and application-learned behaviour. This may include several forms of learning tools, such as real-time virtual/ collaboration software, self-paced web-based courses, electronic performance support systems embedded within the job-task environment, and knowledge management systems. Blended learning often may be a mix of traditional instructor-led training, synchronous online conferencing or training, asynchronous self-paced study, and structured on-the-job training from an experienced worker or mentor

Research plan

Research questions

The primary questions the evaluation was designed to answer are:

1. what effect does blending learning have on measured skills in English compared to face-to-face learning?
2. what effect does blending learning have on measured skills in maths compared to face-to-face learning?

The range of outcome measures to be considered is described in the outcomes section below.

Design

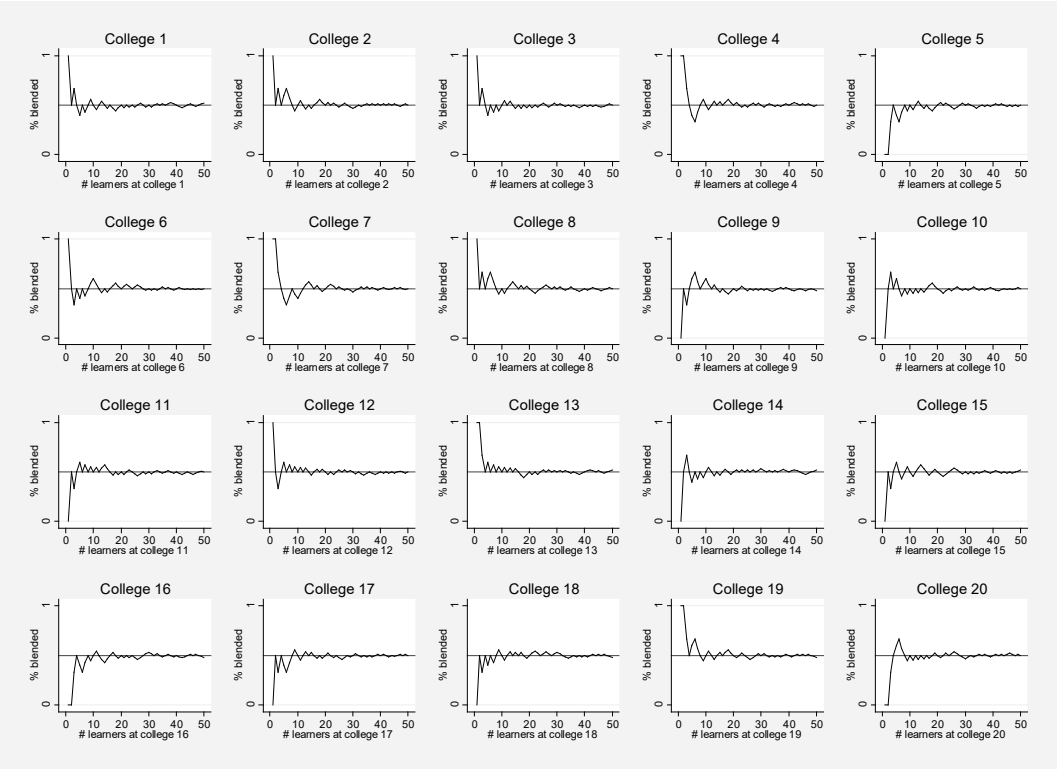
The trial involves two arms; face-to-face learning and blended learning. Randomisation is at the level of the individual learner. Randomisation is carried out within blocks (colleges/centres). Individuals are randomised during course enrolment.

Within each centre, learners' randomisation outcomes will be determined on the basis of a pre-generated randomised sequence. Successive entrants to the trial are given the next available outcome for the sequence. There are separate sequences for English and maths.
Randomisation sequences have been generated using permuted block randomisation with a 50:50 allocation ratio. The procedure is as follows:

- A dataset with 4 observations was created.
- An "arm" variable was created; two observations had arm=”F” and two had arm=”B” (face-to-face and blended, respectively).
- These four observations were then duplicated so that there were now 10,000 observations.

- A "block" variable was created that identified each block (i.e. running from 1 to 2500).
- A random number was created.
- The dataset was sorted by block and, within block, by the random number.
- The resulting sequence of B's and F's is the randomisation sequence.

This approach avoids randomisation outcomes being predictable, yet still converges quickly to the required 50:50 allocation. The chart below shows this for 20 sequences (a separate sequence is used for each course within each centre).



Randomisation will be monitored to ensure that individuals are allocated in line with the randomisation sequence for their centre and course. As the dataset of randomised individuals grows, the extent to which background characteristics (including the pre-test) are similar across the two arms of the trial will be monitored.

Participants

The trial involves working closely with centres, so only learners within participating centres are eligible. Eligible learners are those enrolling on English or maths courses at any level from Entry Level 1 up to Level 2. Learners participate on an informed consent basis. In addition to being willing to participate in the trial, there must be both a face-to-face class and a blended class that the learner is able to attend.

The first step in recruiting centres was to carry out a survey asking about the nature of their provision. Centres that showed interest in being involved in the trial were followed up in order to establish their suitability and, if appropriate, to obtain their agreement to full participation. To be suitable, they must: deliver both face-to-face and blended learning; have sufficient numbers of learners for basic English and maths courses; and be willing to introduce randomisation into their enrolment process.

Once centres have agreed to be involved, randomisation will take place as part of their enrolment process. Project caseworkers will be on hand to assist with this. Appendix 1 gives details of the instruction passed to teachers and administrators. Appendix 2 shows the background questionnaire that is asked of all eligible learners prior to randomisation. Learners are asked their consent to:

- complete the survey (question 1)
- take part in the research and be randomised (question 50)
- take part in a follow-up survey (question 51)
- have their survey answers linked to other information on adult learning (questions 52/54)
- have their survey responses linked to a learner dataset with benefit and employment data (questions 53/55)

A CONSORT diagram will be used to present a summary of the numbers of learners recruited, randomised and, assessed. The number of learners included at each stage will be given, together with an explanation for those excluded.

Outcome Measures

Skill levels will be assessed at the start of each learner's course and again at the end (pre- and post-tests). The assessments allow each individual's skill level for a given subject domain to be put on a scale that is common across skill levels and therefore allows comparability across all learners. The **primary outcome** (for both English and maths learners) is the post-test. For English, three domains will be assessed: reading, writing and speaking, listening and communication. Maths is a single assessment.

Individuals involved in the trial will also be asked to be part of a longitudinal survey and to have their data linked to matched ILR-DWP-HMRC administrative data. **Secondary outcomes** will include longer-term attainment (measured approximately one year after course completion as part of the longitudinal survey and, using tests that resemble as closely as possible the pre- and post-tests described above). Also, changes in confidence in English/maths, labour market outcomes (employment, earnings, benefits receipt) taken from both the longitudinal survey and the administrative data and subjective well-being.

The nature of the intervention is such that it is not possible to randomise after the pre-test (the process of enrolling to courses would become too lengthy). Instead, the 'pre-tests' will be carried out as early as possible into each learner's course. The concern with this is that test results may be influenced by the mode of learning. However, the early weeks of the course are concerned mainly with diagnostic assessments rather than formal tuition so it is unlikely that the 'pre-tests' will be compromised. Both the pre- and post-tests will be computer-delivered, which limits the scope for any systematic differences across the arms of the trial in the administration of the tests to be introduced.

Sample size calculations

The target numbers for the trials are 750 English learners and 750 maths learners. We assume that focusing on the change in attainment rather than post-test attainment *per se* will reduce the variance of the outcomes by 30 per cent. Requiring 2-tail tests with 80% power and 95% significance implies, for each trial, a minimum detectable effect size of 0.17.

Analysis plan

Analysis will be carried out separately for English learners and maths learners. All randomised learners will be included and will retain their randomisation outcome, regardless of whether they subsequently drop out or change from one arm to another. Hence, the impact estimate will capture intent to treat (ITT), Analysis will use linear regression with robust standard errors. Both individual-level and centre-level regressors will be included. Centre identifiers will be included among the regressors to account for the stratification of randomisation. Where outcome data are missing, observations may have to be dropped – this will be fully reported. We will also explore whether bounds analysis can give informative results.

A number of subgroups will be considered. Our ability to do so will depend on the composition of the achieved sample. However, subgroups of particular interest include those defined on the basis of:

- sex
- age
- qualification level
- prior experience of functional skills training
- employment status
- level of IT literacy
- ESOL
- ethnicity
- implementation characteristics.

While the evaluation does not include a separate process study, caseworkers will report on activities in centres pertinent to trial results, and ensure that detailed practical

information is captured on the extent to which the planned trial design has been possible in terms of centres delivering programmes. The caseworker team will be supported by a central project office, staffed by AlphaPlus and NIACE staff, who will ensure caseworkers are provided with the resources they need. The project office is responsible for ensuring the trial activities are compatible with the project design provided by NIESR.

Personnel

- NIESR: Richard Dorsett and Cinzia Rienzo
- AlphaPlus: Jenny Smith
- NIACE: Sue Southwood
- The responsibilities of these organisations are as follows:
  - NIESR is responsible for: the design of the trial; monitoring of randomisation; analysis and reporting of trial.
  - AlphaPlus and NIACE are jointly responsible for recruiting centres and implementing the trial in line with the design.

Timeline

- The timetable is dictated by how quickly the required sample size can be achieved, and also by how long learners' courses tend to be. The current expectation is shown below:

  - Intake to the trial will begin in September 2013 and is likely to continue until at least January/February 2014 (AlphaPlus)
  - Impacts using post-tests reported December 2014 (NIESR)
  - Impacts using matched administrative data and survey data by September 2016 (NIESR)

Risks

Some of the key risks are listed below:

- Should the level of volunteering to participate in the trial be low, there are two consequences. First, it will take longer for the target number of participants to be achieved, so the timetable will slip. Second, it will raise concerns about the generalisability of the results. The likely level of volunteering is unknown *a priori* but will need to be closely monitored, particularly in the early days of randomisation.
- Some participants in the trial will drop out or will not provide outcome data. While the delivery team will be focused on minimising the extent of this, where it does occur it can create analytical problems. The extent of attrition will be reported. Analytical techniques will be used to examine whether it is likely to bias impact estimates. Similarly, there may be instances where learners withdraw from the

research and request that their data not be used. In such cases, there is little option but to exclude observations for that individual. Alternatively, a teacher could decide to withdraw a learner or group of learners. This is perhaps unlikely but could arise if, for instance, learners were seen to be doing far better on one arm than the other.

- In some cases, fidelity may lapse. This creates difficulties in interpreting impact estimates. Caseworkers will be in regular contact with centres to understand the nature of the learning provided under each of the arms. They will aim to ensure that the two arms in the trial deliver learning consistent with the adopted definitions of face-to-face and blended.

B. Characteristics of learners

B.1 Characteristics of learners for whom we have pre-tests

This section presents the results of the balancing analysis for the sample of learners who completed pre-tests.

The average age of participants that completed the pre-test was approximately 35 years of age (Table B.1), essentially the same as in the broader sample of randomised individuals. There is no indication of statistically significant differences between the treatment and control groups in the reading, writing or maths samples.

Table B.1 Age at start of study (years)

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	35.16	34.24	34.65
N	63	79	142
		t= 0.51	p= 0.61
Writing sample			
Mean	34.69	34.33	34.49
N	61	76	137
		t= 0.19	p= 0.85
Maths sample			
Mean	36.04	34.43	35.19
N	77	86	163
		t= 0.87	p= 0.39

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

The proportion of participants that were female (Table B.2) is even higher than it was among the randomised sample, implying that male participants were less likely to go on to complete their pre-test. Again, there is no evidence of a statistically significant difference in this characteristic opening up between the treatment and control groups.

Table B.2 Proportion of sample female

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	0.83	0.77	0.8
N	63	79	142
		t= 0.78	p= 0.44
Writing sample			
Mean	0.84	0.76	0.8
N	61	76	137
		t= 1.05	p= 0.30
Maths sample			
Mean	0.77	0.79	0.78
N	78	86	164
		t= -0.33	p= 0.74

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

On average, members of the reading and writing sample left full time education at the age of 20 (Table B.3), while members of the maths sample on average left closer to age 18. There are only very small differences in this characteristic between the blended learning and face-to-face arms, and these were not statistically significant.

Table B.3 Age left full time education

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	19.93	20.03	19.99
N	43	61	104
		t= -0.09	p= 0.93
Writing sample			
Mean	20.21	19.87	20.02
N	42	55	97
		t= 0.28	p= 0.78
Maths sample			
Mean	18.18	18.39	18.29
N	65	70	135
		t= -0.26	p= 0.80

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

There is no evidence of any difference in the ethnic group distribution (Table B.4) between the treatment and control arms in reading, writing or maths samples. However, the sample has become less representative of the national population, with learners from ethnic minorities more over-represented than was the case in the randomised sample. Again, this is likely explained by the large proportion of the sample based at one large centre (as mentioned in Section 5.1.2) , which also had a lower attrition rate than some other colleges.

Table B.4 Ethnicity, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
White	33.3	24.1	28.2
Mixed	4.8	3.8	4.2
Asian	15.9	21.5	19
Black	33.3	39.2	36.6
Other	11.1	8.9	9.9
Prefer not to say	1.6	2.5	2.1
Total	100	100	100
N	63	79	142
	Pearson chi2(5) = 2.3989 Pr = 0.792		
Writing sample			
White	34.4	26.3	29.9
Mixed	3.3	5.3	4.4
Asian	14.8	21.1	18.2
Black	32.8	34.2	33.6
Other	13.1	10.5	11.7
Prefer not to say	1.6	2.6	2.2
Total	100	100	100
N	61	76	137
	Pearson chi2(5) = 2.1504 Pr = 0.828		
Maths sample			
White	41	46.5	43.9
Mixed	3.8	4.7	4.3
Asian	15.4	14	14.6
Black	29.5	24.4	26.8
Other	7.7	9.3	8.5
Prefer not to say	2.6	1.2	1.8
Total	100	100	100
N	78	86	164
	Pearson chi2(5) = 1.3547 Pr = 0.929		

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

The partnership statuses of participants who have taken the pre-test (Table B.5) are well-balanced within the treatment and control arms. In the case of reading and writing, there is something of a decrease in the proportion of those in relationships, compared to that

found in the overall randomised sample. In addition, there is a higher rate of maths learners being partnered than is the case for the reading or writing samples.

Table B.5 Partnership status, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
Single	61.9	57	59.2
Partnered	34.9	34.2	34.5
Prefer not to say	3.2	8.9	6.3
Total	100	100	100
N	63	79	142
			Pearson chi2(2) = 1.9383 Pr = 0.379
Writing sample			
Single	63.9	57.9	60.6
Partnered	34.4	32.9	33.6
Prefer not to say	1.6	9.2	5.8
Total	100	100	100
N	61	76	137
			Pearson chi2(2) = 3.5492 Pr = 0.170
Maths sample			
Single	51.3	55.8	53.7
Partnered	41	38.4	39.6
Prefer not to say	7.7	5.8	6.7
Total	100	100	100
N	78	86	164
			Pearson chi2(2) = 0.4444 Pr = 0.801

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

The economic activity of participants (Table B.6) does not change dramatically between the complete randomised sample and those taking the pre-test. There is a tendency

towards higher proportions of part time workers in the reading and writing samples. The economic activity of workers remains well-balanced across treatment arms.

Table B.6 Economic activity, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
Full time Work	31.7	29.1	30.3
Part time Work	12.7	11.4	12
Unemployed	30.2	38	34.5
Economically Inactive	9.5	12.7	11.3
Other	3.2	2.5	2.8
Prefer not to say	100	100	100
Total	63	79	142
N			
			Pearson chi2(5) = 2.6608 Pr = 0.752
Writing sample			
Full time Work	9.8	6.6	8
Part time Work	39.3	27.6	32.8
Unemployed	9.8	11.8	10.9
Economically Inactive	27.9	36.8	32.8
Other	9.8	13.2	11.7
Prefer not to say	3.3	3.9	3.6
Total	100	100	100
N	61	76	137
	Pearson chi2(5) = 3.1755 Pr = 0.673		
Maths sample			
Full time Work	11.5	11.6	11.6
Part time Work	25.6	27.9	26.8
Unemployed	6.4	5.8	6.1
Economically Inactive	35.9	31.4	33.5
Other	16.7	17.4	17.1
Prefer not to say	3.8	5.8	4.9
Total	100	100	100
N	78	86	164
			Pearson chi2(5) = 0.6887 Pr = 0.984

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

B.2 Characteristics of learners for whom we have pre-tests and post-tests

This section presents the results of the balancing analysis for the sample of learners who completed both pre-tests and post-tests.

The English samples are broadly in line with the average age seen in the sample that was randomised, while the maths sample is approximately two years older, on average, than the randomised sample (Table B.7). However, these are not dramatic changes in the sample characteristics. There is almost a five year age difference between the face-to-face and blended learning arms, although this is not quite statistically significant at the 10% level.

Table B.7 Age at start of study (years)

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	36.58	34.33	35.27
N	31	43	74
		t= 0.86	p= 0.39
Writing sample			
Mean	38.25	33.4	35.74
N	28	30	58
		t= 1.64	p= 0.11
Maths sample			
Mean	37.87	34.89	36.44
N	39	36	75
		t= 1.10	p= 0.28

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

The proportion of participants that were female (Table B.8) is higher than it was for the sample of randomised participants, implying that male participants were less likely to go on to complete pre- and post-tests. The proportion that were female differs rather substantially between treatment and control arms in the reading and writing post-tested samples, but these are not statistically significant differences.

Table B.8 Proportion of sample female

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	0.84	0.74	0.78
N	31	43	74
		t= 0.97	p= 0.34
Writing sample			
Mean	0.86	0.70	0.78
N	28	30	58
		t= 1.43	p= 0.16
Maths sample			
Mean	0.82	0.78	0.8
N	39	36	75
		t= 0.46	p= 0.65

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

The average age that participants reported having left full time education has risen still further than its level for the randomised sample and the sample completing pre-tests (Table B.9). Most dramatically, the average age for those in the writing sample is now consistent with the average participant having graduated from higher education; the two English samples are notably higher than those for maths, perhaps reflecting individuals from overseas with relatively high levels of education but weak English skills. The increase in the average age relative to the randomised sample also may also highlight an increased risk of non-completion among those who left school earlier.

Table B.9 Age left full time education

	Face-to-face	Blended learning	Overall
Reading sample			
Mean	19.35	20.61	20.12
N	20	31	51
		t= -0.77	p= 0.44
Writing sample			
Mean	21.39	21.35	21.37
N	18	20	38
		t= 0.02	p= 0.98
Maths sample			
Mean	18.67	18.67	18.67
N	33	30	63
		t= 0.00	p= 1.00

Notes. Reporting mean characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. t and p values report the results of a statistical significance test of the null hypothesis of no difference between the means for Face-to-face and Blended Learning groups. Individuals may appear in multiple sections of the table.

The ethnicity of the post-tested sample remains balanced across treatment arms (Table B.10). It is, however, notable that in the reading and writing samples individuals from a Black ethnic background are now the largest single group, replacing those from a White ethnic background, who were the largest single group in the randomised sample. In the case of maths, both Black and White have become proportionally larger groups, while there is a fall in the share of the sample from Asian backgrounds.

Table B.10 Ethnicity, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
White	29	11.6	18.9
Mixed	3.2	4.7	4.1
Asian	12.9	27.9	21.6
Black	41.9	46.5	44.6
Other	12.9	4.7	8.1
Prefer not to say	0	4.7	2.7
Total	100	100	100
N	31	43	74
	Pearson chi2(5) = 7.8892 Pr = 0.162		
Writing sample			
White	25	26.7	25.9
Mixed	3.6	6.7	5.2
Asian	10.7	20	15.5
Black	42.9	33.3	37.9
Other	17.9	10	13.8
Prefer not to say	0	3.3	1.7
Total	100	100	100
N	28	30	58
	Pearson chi2(5) = 3.0164 Pr = 0.697		
Maths sample			
White	41	38.9	40
Mixed	0	8.3	4
Asian	5.1	13.9	9.3
Black	41	30.6	36
Other	10.3	8.3	9.3
Prefer not to say	2.6	0	1.3
Total	100	100	100
N	39	36	75
	Pearson chi2(5) = 6.3780 Pr = 0.271		

Notes. Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

Turning to partnership status (Table B.11), in the reading and writing post-tested samples the single group have increased their shares relative to those seen in the randomised sample. Conversely, the opposite is the case in the maths sample. Nevertheless, in all

three samples there remain no significant differences in partnership status between face-to-face and blended learning arms.

Table B.11 Partnership status, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
Single	67.7	58.1	62.2
Partnered	32.3	37.2	35.1
Prefer not to say	0	4.7	2.7
Total	100	100	100
N	31	43	74
			Pearson chi2(2) = 1.8347 Pr = 0.400
Writing sample			
Single	60.7	60	60.3
Partnered	35.7	30	32.8
Prefer not to say	3.6	10	6.9
Total	100	100	100
N	28	30	58
			Pearson chi2(2) = 1.0134 Pr = 0.602
Maths sample			
Single	46.2	55.6	50.7
Partnered	48.7	44.4	46.7
Prefer not to say	5.1	0	2.7
Total	100	100	100
N	39	36	75
			Pearson chi2(2) = 2.2460 Pr = 0.325

Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

Finally, we consider economic activity (Table B.12). There are substantial changes in the distribution of economic activity relative to the randomised sample, with those in part time work being rather more likely than average to complete a post-test, while those who are

economically inactive or in full time work are less likely to do so. Despite this change in the average characteristics of the group, the characteristic remains balanced across treatment and control groups.

Table B.12 Economic activity, column percentages

	Face-to-face	Blended learning	Overall
Reading sample			
Full time Work	9.7	9.3	9.5
Part time Work	48.4	32.6	39.2
Unemployed	3.2	9.3	6.8
Economically Inactive	29	30.2	29.7
Other	6.5	18.6	13.5
Prefer not to say	3.2	0	1.4
Total	100	100	100
N	31	43	74
			Pearson chi2(5) = 5.5034 Pr = 0.358
Writing sample			
Full time Work	7.1	10	8.6
Part time Work	60.7	33.3	46.6
Unemployed	3.6	3.3	3.4
Economically Inactive	21.4	30	25.9
Other	3.6	20	12.1
Prefer not to say	3.6	3.3	3.4
Total	100	100	100
N	28	30	58
		Pearson chi2(5) = 6.1246 Pr = 0.294	
Maths sample			
Full time Work	12.8	11.1	12
Part time Work	30.8	44.4	37.3
Unemployed	5.1	5.6	5.3
Economically Inactive	30.8	27.8	29.3
Other	15.4	11.1	13.3
Prefer not to say	5.1	0	2.7
Total	100	100	100
N	39	36	75
			Pearson chi2(5) = 3.1494 Pr = 0.677

Reporting column percentages of characteristics by treatment arm for all successfully randomised participants who complete relevant pre- and post-tests. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between the distribution of the characteristics and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

We should note, however, that Table B.13 highlights the high rate of attrition we have seen overall between randomisation and post-test in this trial. There is no evidence of statistically significantly different rates of attrition by treatment arm, although the difference is not small (13% vs. 18%) in the case of the reading sample.

Table B.13 Attrition between randomisation and post-test, row percentages

	Attrition	Completion	Total	N
Reading sample				
Face-to-face	86.8	13.2	100.0	235
Blended Learning	81.9	18.1	100.0	237
Overall	84.3	15.7	100.0	472
	Pearson chi2(1) = 2.1888 Pr = 0.139			
Writing sample				
Face-to-face	88.1	11.9	100.0	*235*
Blended Learning	87.3	12.7	100.0	*237*
Total	87.7	12.3	100.0	*472*
	Pearson chi2(1) = 0.0605 Pr = 0.806			
Maths sample				
Face-to-face	80.8	19.2	100.0	*203*
Blended Learning	80.9	19.1	100.0	*188*
Total	80.8	19.2	100.0	*391*
	Pearson chi2(1) = 0.0002 Pr = 0.987			

Reporting row percentages of pre- and post-test completion by treatment arm for all successfully randomised participants. chi2 and p values report the results of a statistical significance test of the null hypothesis of no association between drop-out and being in Face-to-face or Blended Learning groups. Individuals may appear in multiple sections of the table.

D. Calculation of effect sizes

The effect size, g, is calculated as:

$$g = \frac{\bar{x}_1 - \bar{x}_2}{s^*}$$

where s* is the pooled standard deviation, calculated as:

$$s^* = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

However, this is known to be biased. This bias can be overcome by applying a correction factor J:

$$J(a) = \frac{\Gamma(a/2)}{\sqrt{a/2}\,\Gamma((a-1)/2)}$$

where Γ is the gamma function. The adjusted estimator, g*, becomes:

$$g^* = J(n_1 + n_2 - 2)\, g$$

In practice, this is often unfeasibly computationally intensive due to the explosive nature of the gamma function. We follow standard practice and use an approximation to give the bias-corrected Hedges' g*:

$$g^* \approx \left(1 - \frac{3}{4(n_1 + n_2) - 9}\right) g$$

All effect sizes in the tables in the main body of the report use this formula. As such, they provide an estimate of the effect of blended learning relative to face-to-face teaching in units that are comparable across outcome measures and with other trials in the literature.

Department for Education

Reference: DFE-RR794

ISBN: 978-1-78105-874-9

This research was commissioned under the 2010 to 2015 Conservative and Liberal Democrat coalition government. As a result the content may not reflect current Government policy. The views expressed in this report are the authors' and do not necessarily reflect those of the Department for Education.

Any enquiries regarding this publication should be sent to us at: vikki.mcauley@education.gov.uk or www.education.gov.uk/contactus

This document is available for download at www.gov.uk/government/publications