Early Years Toolbox

Pilot Report

May 2020

Anneka Dawson, Lucy Stokes, Clare Huxley, Johnny Runge, Helena Takala, Chiara Manzoni, Nathan Hudson- Sharp and Ceri Williams.

The Education Endowment Foundation (EEF) is an independent grant-making charity dedicated to breaking the link between family income and educational achievement, ensuring that children from all backgrounds can fulfil their potential and make the most of their talents.

The EEF aims to raise the attainment of children facing disadvantage by:

- identifying promising educational innovations that address the needs of disadvantaged children in primary and secondary schools in England;
- evaluating these innovations to extend and secure the evidence on what works and can be made to work at scale; and
- encouraging schools, government, charities, and others to apply evidence and adopt innovations found to be effective.

The EEF was established in 2011 by the Sutton Trust as lead charity in partnership with Impetus (formerly Impetus Trust) and received a founding £125m grant from the Department for Education.
Together, the EEF and Sutton Trust are the government-designated What Works Centre for improving education outcomes for school-aged children.

For more information about the EEF or this report please contact:

Jonathan Kay
Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
SW1P 4QP

0207 802 1653

jonathan.kay@eefoundation.org.uk

www.educationendowmentfoundation.org.uk

Contents

About the evaluator

The project was independently evaluated by a team from the Institute for Employment Studies (IES) and the National Institute of Economic and Social Research (NIESR):

Anneka Dawson, Clare Huxley, Helena Takala, and Ceri Williams from IES

Lucy Stokes, Johnny Runge, Chiara Manzoni, and Nathan Hudson- Sharp from NIESR

The lead evaluators were Anneka Dawson and Lucy Stokes.

Contact details:

Name: Anneka Dawson
Address: IES, City Gate, 185 Dyke Road, Brighton, BN3 1TL, UK
Tel: +44 1273 763 445
Email: Anneka.Dawson@employment-studies.co.uk

Executive summary

The project

The Early Years Toolbox (EYT) is a set of iPad-based assessment tools suitable for use with young children by early years staff. It includes eight brief, engaging, and game-like assessment apps that aim to capture abilities that research has shown to be the most predictive of later academic, social, emotional, cognitive, and life outcomes. There is evidence that knowledge of children's skills enables more effective instruction related to individual needs. However, in the early years much less is known about how formative assessment, and knowledge of child development and skills, influences practice and outcomes. In this pilot trial, Action for Children, supported by Professor Edward Melhuish from the University of Oxford, provided professional development (PD) to train early years staff to use three of the EYT apps: Expressive Vocabulary, Early Numeracy, and Child Self-Regulation and Behaviour Questionnaire. This pilot trial was co-funded by the Department for Education as part of our Early Years Professional Development Funding round.

The one-year pilot PD programme consisted of an initial training day in regional locations for two staff per nursery, and a two-hour follow up training session held within each nursery for all staff to attend. The nurseries committed to using the apps to assess children once every term and they received a bespoke support visit, timed for after each assessment point, which was tailored to their nurseries' needs. The pilot programme also included an online forum for nursery staff to ask questions or share experiences, a digital pack of evidence-informed resources for language, numeracy, and self-regulation development taken from research literature, and a consolidation day at the end of the year for all nurseries to share best practice.

How was the pilot conducted?

Observations were carried out of the initial training day, the follow-up training, the support visits, and the consolidation event. In addition, all settings were interviewed, either during a case study visit or by telephone interview. An online survey was sent to all practitioners and setting managers before the training and towards the end of the intervention, as well as feedback questionnaires administered by the delivery team at the initial training and consolidation event. A series of three Intervention Delivery and Evaluation Analysis (IDEA) workshops were conducted with the delivery team to help construct and test a theory of change model for the intervention. Analysis of the data collected through the apps was also undertaken with a focus on issues related to feasibility of use, such as time required to play the apps.

The pilot took place over the academic year 2018/2019 in 27 nurseries across three regions in England (South, North-East, and Midlands) with children aged three to four years old. The nurseries were a mix of private, voluntary, independent (PVI), local authority maintained, and school-based settings.

Overview of findings

Summary of pilot findings

Question	Finding	Comment
Is there evidence to support the theory of change?	***Mixed***	***There is some evidence that the intervention improved practice and changed practitioners' behaviour, but only limited evidence on whether the intervention enhanced practitioner understanding of child development, ability to target support, and whether it could improve child outcomes. However, this evidence is based upon staff report that may under or overestimate change.***
Was the approach feasible?	***Yes***	***The approach is feasible and could be improved by automating the presentation of assessment data and making other improvements to gameplay to help increase take-up.***
Is the approach ready to be evaluated in a trial?	***Mixed***	***EYT is now more replicable and scalable but consistency in the cascading of training and consultant delivery is needed to ensure a future trial has the best chance of showing impact.***

		More support is required so that practitioners know how to respond after gaps in children's competencies are identified.

There was mixed evidence found to support the intervention's theory of change. Most practitioners felt that EYT had influenced their practice to some extent, with most suggesting this had been achieved by using the app results to inform planning. This evidences that the intervention fulfilled two of its short-term aims: changing practitioner's behaviour and using knowledge gained from the assessment and training to plan learning experiences. Those that reported no, or limited, change to their practice were typically school-based nurseries that were more satisfied with existing methods of assessment.

However, practitioners in general identified limited change in their understanding of children's development of skills and children's current competencies. Some settings, mostly PVIs (private, voluntary, and independent nurseries), reported greater impacts and suggested that the numeracy app, which presented results in subcategories, was most helpful. There was only limited evidence of EYT data being used to personalise learning; rather, settings tended to introduce whole-class approaches to address the gaps highlighted by the assessments.
Further to this there was limited evidence on whether EYT could achieve its longer-term aim of improving children's attainment. A few practitioners commented on potential changes in children's vocabulary and numeracy, but others reported no perceived impact on attainment as yet. Practitioners felt that training, support visits, and resources need to emphasise how the results from the apps can be used to inform practice and provide information on how to tailor support to children in particular need.

Overall, the approach used for the pilot of EYT was found to be feasible, but that it might be worth revisiting the approach to cascading training to other practitioners at the setting as there were marked differences in the amount of information they received compared to directly-trained practitioners. Additional development of the apps could also be beneficial. Practitioners were positive about the initial training, follow-up training, the support visits, and the consolidation day. However, practitioners who had not received the initial training were less likely to understand the full intervention beyond child assessment through the apps because of limited cascading practices by practitioners who attended the training day. The phone, email, and online forum support were not widely used, although this was typically because settings did not feel the need for additional support. Further to this, the additional evidence-based resources for language, numeracy, and self-regulation development were not well used, with some practitioners not aware of them.

On average, settings did not use the apps as often as intended. Most settings conducted the assessments twice (rather than three times), although issues in the recording of app data may mean that the frequency of app usage was underestimated. Time to complete assessments was the most reported barrier to implementation, with many practitioners feeling the self-regulation app was particularly time consuming. At present there is no automated way of collecting assessment results from the apps in a form that would be user-friendly for practitioners and some practitioners found the current process too effortful. A few further issues with the apps themselves were identified, for example, children experiencing difficulties counting particular images, which caused some concerns among practitioners about the accuracy of the data collected.

To improve the scalability of the intervention, more guidance is needed on cascading to other practitioners in a setting. Consultants who deliver the support visits also need more consistent training. Ideally the EYT intervention could be enhanced by more support and signposting to specific resources after practitioners have discovered gaps or low performance in the app-based assessments as well as improvements to the apps themselves by automating recording of assessment and reducing game-play issues.

Introduction

Intervention

A description of the intervention is provided below using the Template for Intervention Description and Replication (TIDieR) framework as created by Hoffman et al. (2014) and adapted by the EEF.

1. Brief name

The Early Years Toolbox (EYT): piloting the use of an objective child assessment method with early years workers

2. Why (rationale/theory)

Action for Children (AfC) proposed to introduce an app-based method to enable early years practitioners to have objective knowledge of children's competencies and skills. Howard and Melhuish (2015) demonstrated that the apps worked well with trained assessors, however, there are no evaluations of how the apps might be integrated into routine practice by early years practitioners, or whether they have a beneficial effect upon their practice and interactions with children. By providing staff with information about children's emerging abilities using the EYT, the idea is that practitioners will be better placed to ensure children's experiences in the settings are tailored to their individual development needs and support visits will guide practitioners in how to use this information about children's abilities. This is different from business as usual in the settings where using assessments with this age group was mixed in respect to frequency and quality of measurement. Furthermore, the pilot seeks to establish a viable, scalable programme that could then be rolled-out as part of a larger scale RCT that thoroughly tests the assumptions and hypotheses underlying this approach.

3. Who (recipients)

Early years practitioners working with children aged three to four at 27 participating early years settings. These included ten primary schools, four maintained nursery schools, and 13 PVI providers. Six of the settings were Action for Children nurseries. Settings were expected to use the EYT apps with all three-year-old children in the setting, at least once per term, where possible.[1]

4. What (materials)

Settings were provided with iPads and were given the Early Years Toolbox app package for use with the iPads. The Early Years Toolbox is a series of apps designed to measure young children's emerging cognitive, self-regulatory, language, and social development (example screenshots can be seen in Appendix 6). Each measure is a brief, game-like assessment that has been developed for the iPad. Developed by early years researchers Professor Edward Melhuish (University of Oxford) and Dr Steven Howard (University of Wollongong, Australia), the EYT is designed to capture abilities that research has shown to be most predictive of later academic, social, emotional, cognitive, and life outcomes. For the purposes of this project, three apps were used: a numeracy app for the children to complete with a practitioner ('EYT Early Numeracy'), a vocabulary app for the children to complete with a practitioner, and a self-regulation and social development app for the practitioner to complete alone or in pairs about the children (referred to as 'self-regulation' for brevity after this point). The EYT also includes several other apps but these were not used in this project. The three apps were selected for the pilot by the delivery team as these apps are considered to be those most closely related to the early years curriculum and the Early Years Foundation Stage Profile. In addition, AfC also provided practitioners with a digital pack of evidence-based materials on numeracy, language development, self-regulation, and social development taken from the research literature and a short summary document with suggestions for the different areas and further reading at the follow-up training session. Further resources focused on social development and self-regulation were shared in the spring term. Additional resources were also shared or signposted by the consultants in the support sessions dependent on the nurseries' needs in these areas.

[1] Including children who were aged three at the start of the study but who turned four during the life of the project. In some larger settings, it was agreed that the app would only be used with one class or room.

5. What (procedures)

The Toolbox intervention included the EYT with wraparound support. Firstly, key staff from the participating settings attended an initial regional training day in the first half-term of the academic year delivered by Professor Edward Melhuish and AfC. The training covered how to use the three apps, which are the focus of this project, and the background to why these three areas are important to child development and their relation to longer term outcomes. The practitioners then had the opportunity to practice using the apps with some children from the setting they were being trained at and ask any questions. Staff were then expected to cascade the training to other staff in their setting working with three-year-olds using the slides from the training. Practitioners had flexibility to do the cascading to other staff as was convenient to them.

Each setting then received a follow-up training visit from a member of the AfC team (called a consultant) of approximately two hours before Christmas (after about two months of delivery). The follow-up session gave practitioners the chance to discuss how cascading had gone, gave the consultant the opportunity to observe practitioners carrying out the apps with children, and then for practitioners to ask any questions they had including asking for specific resources to support their setting's needs.

Once the initial and follow-up training was completed, there was an ongoing package of support provided through the delivery period:

- Settings were visited by their consultant for support visits, with most settings receiving a total of three of these visits over the course of the intervention (three settings started later due to some technical issues with using their own iPads and only had two visits). The support visits were all about two hours in length and included time for questions about issues specific to the setting. For example, settings could ask for specific advice or resources about an area they had identified through the apps as something they wanted to focus on.
- A consolidation event was held in the summer term before the last support visit, which shared learning across teams and covered how to manage data and how to use it for school transitions as well as practical activities on how to implement numeracy into art lessons.
- Further support could be accessed via email and telephone if required.
- An online forum was established in early 2019 in order to provide further support and to give the opportunity for participants to share learning and progress.
- AfC provided practitioners with a digital pack of evidence-based materials as outlined above.

6. Who (implementers)

The intervention was delivered to early years practitioners by AfC and Professor Melhuish. Early years practitioners used the EYT apps with participating children in their setting.

7. How (mode of delivery)

The EYT numeracy and literacy apps were designed to be played as games with individual children as part of everyday activities rather than being perceived as a formal assessment. The self-regulation and social development app involved forms to be completed on the iPad outside of the class environment by the practitioner who was most familiar with the child.

8. Where (setting)

The EYT apps were intended to be used with children in a suitable location at the setting (no internet connection was required to use the app). The settings were located in three areas in England: North, South, and West Midlands.

9. When and how much (dosage)

The intervention was carried out in settings from September 2018 to July 2019. EYT assessments were carried out with all participating children once a term where possible. Practitioners at settings received three support visits from AfC consultants between January and July 2019, as well as ongoing access to other modes of support.

10. Tailoring

The EYT apps are standardised measures so assessments needed to be delivered as specified and consistently with each child. Some practitioners chose to use the self-regulation app in pairs and others individually. AfC did not initially suggest this to practitioners, but once it knew some practitioners had been doing it in pairs it encouraged other settings to do the same However, delivery team consultants were able to adapt support sessions for practitioners in response to any needs identified through using the app. This included providing resources for teaching with children as well as training and resources for practitioners.

11. How well (planned): strategies to maximise effective implementation

Strategies to maximise implementation effectiveness included one full day of initial training with practitioners in October and a two-hour follow-up training visit at the setting before Christmas. This was supplemented by an ongoing support package, including three support visits from AfC consultants.

Background evidence

The EYT is a simple to use, quick set of measures and has shown promising results in the Australian context where it has demonstrated good internal consistency and convergent validity with other early years' measures (Howard and Melhuish, 2016). It has been used as an outcome measure in several evaluations (for example, the Study of Early Education and Development, SEED study, Natcen, no date). The EYT is widely used in Australia and is currently being used across five continents (Howard, Melhuish and Chadwick, 2019) in a range of settings in both education and research across a variety of outcomes. However, there have been no evaluations exploring the EYT's role on professional practice or child outcomes, or how the package of overall support making up the Toolbox intervention being explored in this pilot may enhance the apps' use.

In research with school teachers there is evidence that knowledge of children's skills enables more effective instruction related to individual needs (Connor, Morrison, Fishman, Schatschneider and Underwood, 2007; Connor et al., 2011; York, 2014). However, in the early years field, the developers felt there was little robust evidence available on how formative assessment, and knowledge of child development and skills, influences practice. Furthermore, more research is needed about how assessment might be universally integrated into routine practice by early years staff, and whether this has a beneficial effect upon educators' practice or children's outcomes.

The Early Years Toolbox therefore was developed to provide early years practitioners with an easy to use, low cost, robust way of measuring children's abilities which can help inform their practice at both the individual pupil and class level. The intervention, which seeks to integrate the EYT into the routine practice of early years practitioners, is at the pilot stage of development. The three EYT apps of relevance to this project have been developed based on research across the three areas of vocabulary, numeracy, and social development and self- regulation. These three areas are all linked to outcomes later in life such as language (Rowe, Raudenbush, and Goldin-Meadow, 2012), social development and social interactions (Vandell et al., 2010), school engagement (Boeckaerts and Corno, 2005), and language and literacy (Reese, Sparks, and Leyva, 2010). There is also research showing that disadvantaged children have a gap in their abilities in these areas before age five, which can have a long term effect on outcomes (Hart and Risley, 1995). The developers believe that being able to monitor early vocabulary, numeracy, and social development and self-regulation will enable early years practitioners to focus in on struggling children at a very early stage before they reach formal schooling and potentially intervene to help improve those children's outcomes. Although other measurement tools were available, the developers felt that many of them had limitations such as not being clear in the construct that is being measured and computerisation limiting the ability for the child to respond to the measure themselves directly, which could pose a barrier. These measures have been developed with young children in mind and using an iPad without the need for continuous internet access means they are more accessible and pragmatic for both children and practitioners. This includes ensuring they are developmentally appropriate in respect to instructions and task requirements, brief to deliver, and developmentally sensitive. The developers established norms, validity, and reliability by analysing the results from 1,764 children who had used the apps.

The evaluation of the EYT pilot seeks to explore the feasibility and acceptability of providing more formative information to early years practitioners through the intervention to see if this influences their practice, and what level of information

and support they need to interpret and act on the assessments as well as whether the intervention could potentially be scaled up to a larger number of early years settings. It also aims to build new evidence on the possible benefits of formative assessment in early years (nursery) such as child outcomes. If the pilot found the approach to be feasible and beneficial, the EEF could consider funding a trial to test the impact on child outcomes.

Research questions

The aim of the evaluation was to examine whether the pilot programme shows evidence of promise, was feasible for providers to deliver, and can be replicated, scaled, and tested as part of a larger efficacy trial. This pilot evaluation aimed to answer the following detailed research questions in each of these areas in order to direct the collection of evidence and show whether these conditions have been met.

Whether there is evidence to support the theory of change

- Is there evidence of an improvement in practitioners' practice and behaviour to better support children's early development?
- Is there evidence of an improvement in practitioners' understanding of pupils' competencies and skills?
- Do participants perceive a change in work practices from engaging in the intervention?
- Do participants believe that the intervention has had (or will have) benefits to children compared to previous practices?
- Are there any unintended positive or negative effects of the intervention?
- Are practitioners more confident that they are correctly targeting children in need of support?

Whether the approach is feasible to deliver

- Is the intervention feasible for practitioners to engage in and implement?
- Can the intervention be delivered within routine practice?
- What are the facilitators/barriers to the programme?
- What level of support is needed to enable practitioners to effectively interpret assessment data?
- What level of support is needed to enable practitioners to effectively change their practice based on this data?
- What is the optimal or most feasible frequency and timing of the assessments within the school term?

Whether the programme is ready to be evaluated in a trial

- Do the training and resources constitute a defined intervention that can be replicated and scaled up?
- What, if anything, should be updated or changed to prepare the intervention for a trial evaluation (for example, theory, materials, procedures)?
- What will the cost of the intervention be to settings (both direct costs of training and indirect costs, for example of interventions bought as a result of the assessments)?

Note that in presenting the findings in this report, in order to aid the flow of the discussion, we structure these around the key themes and topics for each of the three broad areas (and therefore, while all individual research questions listed above are covered in the discussion, findings are not necessarily described under headings of the detailed research questions).

Ethical review and data protection

The project received ethical approval from the NIESR Research Ethics Committee. This required the submission of an application form to the ethics committee outlining the key features of the project and setting out the ethical issues involved and associated mitigations.

Settings were provided with a Memorandum of Understanding (see Appendix 1) explaining what the project entailed and the responsibilities of the evaluators, the delivery team, and participating settings. Parents of eligible children were

sent a letter explaining the study and giving the option for parents to withdraw their children from the study, which they are able to do at any time before the data analysis (see Appendix 2).

The pilot involved the processing of personal data relating to staff and children at participating settings. The legal basis for processing personal data was covered by GDPR Article 6 (1) (f) which states that 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

In order to establish this, the evaluation team carried out a legitimate interest assessment which concluded that we had a legitimate interest in processing the personal data in order to conduct the evaluation, the processing identified was necessary in order to meet this purpose, and there would be minimal impacts of the processing on the individuals involved (and which the evaluation team took steps to ensure). During the period of the evaluation, the delivery team (AfC) and the evaluation team (NIESR and IES) were joint data controllers. No special personal data was processed as part of the evaluation.

A privacy notice was developed for the project (Appendix 3) describing the personal data that would be collected and processed as part of the evaluation and why, the legal basis for processing, retention periods, and the rights of individuals to withdraw their data at any time. A data sharing agreement was also drawn up between the evaluation and delivery teams to specify in detail which data would be shared and the processes for doing so. Further details regarding data sharing processes are available from the authors upon request.

Project team

Delivery team:

- Sarah Read (Action for Children)—helped deliver the initial training, quality assurance at follow-up training and support visits, ran a session at the consolidation event, helped put the materials together, and general project support.
- Fliss James (Action for Children)—helped deliver the initial training, delivered follow-up training and support visits with a third of the settings, ran a session at the consolidation event, put the materials together, and general project support.
- Professor Edward Melhuish (University of Oxford)—developed the apps with Professor Steven Howard and delivered the initial training and a session at the consolidation event.
- Clare Lishman—delivered follow-up training and support visits with a third of the settings and support at the consolidation event.
- Liz Somerville—delivered follow-up training and support visits with a third of the settings and support at the consolidation event.

Evaluation team:

- Anneka Dawson (Co-PI, IES)—project direction, quality assurance, development of materials, led the three IDEA workshops, presenting to the EEF, and reporting.
- Lucy Stokes (Co-PI, NIESR)—project direction, quality assurance, development of materials, supporting the three IDEA workshops, presenting to the EEF, and reporting.
- Johnny Runge (NIESR)—fieldwork, supporting the IDEA workshops, development of materials, analysis, presenting to the EEF, and reporting
- Clare Huxley (IES)—project management, fieldwork, supporting the IDEA workshops, development of materials, online surveys, analysis, presenting to the EEF, and reporting.
- Nathan Hudson-Sharp (NIESR)—fieldwork, analysis, and reporting.
- Chiara Manzoni (NIESR)—fieldwork, analysis, and reporting.
- Helena Takala (IES)—fieldwork, analysis, and reporting.
- Ceri Williams (IES)—fieldwork, development of materials, analysis, and reporting.
- Joanne Doherty (IES)—online survey development.

Methods

Recruitment

Settings were recruited to the pilot through Action for Children. The participating settings included some AfC nurseries (six in total) but also other settings in three broad regions: the South (covering Barnet, Hammersmith, Fulham, Essex, Newham, and Bexley), the North-East (covering North Hull, Gateshead, Northumberland, and Stockton), and the Midlands (Birmingham, Derby, Telford, and Wrekin). The regions will be referred to throughout as A, B and C to ensure anonymity. AfC approached hubs that it had worked with previously to help with recruitment, which were all in areas of deprivation. It approached two hubs and one maintained nursery school that had been involved in partnership work in the south, two hubs based in Northumbria and Hull, one hub based in Telford, and two maintained nursery schools that it had worked with previously in the midlands. The hubs then contacted PVI settings that they thought would be interested in participating. Additionally, one setting it had had no prior relationship with expressed an interest in participating during an Action for Children event.

All children at the settings who were three years old at the start of the study in September 2018 were eligible to take part.

Settings were asked to sign a Memorandum of Understanding (see Appendix 1), which set out the requirements of the study and what it entails. Settings were also provided with an information letter for sending to parents/guardians explaining what the study involves and how their child's data will be used. This gives parents clear instructions as to how their child's data can be withdrawn from the evaluation (see Appendix 2). If a withdrawal request was received, practitioners did not use the app with the child.[2] A privacy notice was also made available at: https://www.niesr.ac.uk/projects/pilot-evaluation-early-years-toolbox and can also be seen in Appendix 3.

Data collection

Drawing on the EEF implementation and process evaluation guidance (Humphrey et al., 2016), the study used a multiphase design, based around a triangulation of mixed methods, to help fully develop and explore a theory of change model and then test the three main EEF pilot research questions as outlined in the Introduction.

An initial theory of change model was provided by the delivery team prior to the start of the pilot. This was revised throughout the course of the pilot in response to feedback from expert stakeholders and through a series of three IDEA workshops. The format of the IDEA workshops and the process of developing the theory of change is covered in more detail in the theory of change section of the Findings. The IDEA workshops took place over the course of ten months (October 2018 to August 2019).

An initial discussion of the theory of change for the intervention took place at the first set up meeting for the pilot study. The theory of change was assessed again in the first IDEA workshop, which took place after the practitioner training and also included a review of the programme materials as well as looking at the feedback from practitioners on the training sessions, as collected by AfC. The final two workshops used Connell and Klem's (2000, cited in Humphrey et al., 2016) four criteria for the assessment of the quality of a theory of change. Is it:

- plausible—is it logical and based on good assumptions?;
- doable—can the settings resource it?;
- testable—what methods can be used to test it?; and
- meaningful—is it worth the time/money?—using external feedback (from expert stakeholders and practitioners as part of the implementation and process evaluation).

The second workshop focused on expert stakeholder feedback to the theory of change and updated it as necessary based around these criteria. The final workshop focused on the practitioner feedback from the interviews and case

[2] Information on the number of withdrawal requests received is not available, however, the indications from the analysis of the app data are that the number of withdrawals were low.

studies and allowed final adaptations to the theory of change model as necessary. This final workshop also provided an opportunity to collect cost data from the delivery team. This feedback loop was designed to enable the intervention to be fully developed over the course of the ten months.

Although there is a pre-existing evidence base regarding the validity of the EYT apps as a measure of key areas of child development, the Toolbox intervention, as applied in this study, has not been evaluated with regards to its impact on professional practice in early years teaching or child outcomes. For this reason, and given the pilot stage of the programme, it was deemed appropriate to use primarily qualitative research methods to evaluate the intervention and keep costs and the burden on settings to a minimum.

The evaluation comprised the following data collection methods to examine the research questions outlined in the findings section:

- case study visits to 12 nurseries to interview practitioners and setting managers; six visits took place midway through the programme and the remaining six towards the end of the programme;
- telephone interviews with key staff at the remaining 15 settings (six of these took place midway through the programme and the remaining nine towards the end);
- observations of the three initial group training sessions;
- observations of three of the individual follow-up training sessions;
- observations of eight of the support visits;
- review of programme materials (including the additional resources given to nurseries, consolidation event materials, and the initial training materials);
- analysis of summary documents provided by AfC about the support it had provided in the support visits and how the intervention was progressing;
- analysis of data collected by the delivery team of practitioner feedback surveys from the initial training and consolidation event;
- online survey (baseline and post-treatment) of setting managers and staff to capture change in practices and behaviours during the programme; and
- analysis of child app data as collected through the Early Years Toolbox app.

The case studies and semi-structured interviews explored the experiences and perspectives of practitioners to assess the extent to which the pilot intervention had led to changes in practice and to explore the perceived impacts, or potential impacts, of the intervention. Surveys were used to capture feedback and contextual information from a wide sample of practitioners across the participating settings (copies of the questionnaires used are provided in Appendix 4).

In addition, the interviews explored whether the intervention was seen as feasible and realistic to implement, and identified barriers to implementation and gathered views on how the intervention might be improved to inform potential future rollout. The observations of training days and support visits supplemented the interviews by increasing the evaluation team's understanding of what the intervention looked like in practice, provided a naturalistic impression of how contextual factors affected implementation, and allowed the research team to experience delivery of support in a range of settings and contexts. It also provided an indication as to whether the intervention was implemented in a similar manner across providers.

A census approach was appropriate given the number of settings (27) for the case study/semi-structured interviews and survey elements. It was intended that all settings that signed up would participate in either a case study visit or a telephone interview. The settings were split into two across the regions and half the settings were contacted to arrange visits in spring and then half were contacted in summer. Settings were able to largely self-select whether they were interviewed in person or by telephone so there may be a slight bias—in that more engaged settings are more likely to be open to a visit—but as all settings were interviewed we did not consider this to be a cause for concern. Settings that dropped out were to be contacted for an interview to explore their reasons for this. However, no settings dropped out during the study so these were converted to additional interviews. Similarly, settings were asked to provide contact emails for all practitioners working with the target year group and baseline and post-treatment surveys were sent to all practitioners to complete.

Case study/interview discussion guides, observation frameworks, and survey questionnaires were designed to explore aspects of the agreed research questions. The theory of change model was used as a framework to identify mechanisms by which change was expected to occur, enabling factors, assumptions around conditions for success, target outputs, and outcomes. The research instruments were also informed by discussions with the delivery team during the IDEA workshops and feedback from experts in the fields of psychology and education regarding the theory of change model and the intervention as a whole.

We had planned to examine the support provided by AfC via the helpline, emails, and the online forum but in collaboration with AfC we decided this was not necessary as there was limited contact from practitioners using these methods and they only involved technical issues during set-up at the start which were all resolved. We had also planned to do telephone interviews with settings that dropped out of the intervention over the course of the year but all settings remained involved so this was not necessary; instead we added extra telephone interviews to cover all of the settings (rather than the 24 planned in the study plan).

All evaluation activities were carried out by NIESR and IES with attendance from the delivery team at AfC and Professor Melhuish at the IDEA workshops.

A total of 31 practitioners and 13 managers were interviewed across the settings in the case studies and telephone interviews of which 25 had attended the initial training and 19 had not. Table 2 summarises the total number of interviews at each of the 27 settings.

Table 2: Setting summary of interview completion frequency and format

Identifier	Region	Interviewees	Format
Setting 1	***A***	***1***	***Phone***
Setting 2	***A***	***3***	***Case study***
Setting 3	***A***	***4***	***Case study***
Setting 4	***A***	***1***	***Phone***
Setting 5	***B***	***1***	***Phone***
Setting 6	***B***	***3***	***Case study***
Setting 7	***C***	***3***	***Case study***
Setting 8	***C***	***1***	***Phone***
Setting 9	***B***	***3***	***Case study***
Setting 10	***C***	***1***	***Phone***
Setting 11	***C***	***1***	***Phone***
Setting 12	***C***	***3***	***Case study***
Setting 13	***B***	***1***	***Phone***
Setting 14	***C***	***2***	***Case study***
Setting 15	***B***	***1***	***Case study***
Setting 16	***C***	***3***	***Case study***
Setting 17	***C***	***1***	***Phone***
Setting 18	***C***	***1***	***Phone***
Setting 19	***B***	***1***	***Phone***
Setting 20	***B***	***1***	***Phone***
Setting 21	***B***	***1***	***Phone***

Setting 22	***B***	***2***	***Case study***
Setting 23	***A***	***1***	***Phone***
Setting 24	***A***	***1***	***Phone***
Setting 25	***A***	***1***	***Phone***
Setting 26	***A***	***1***	***Case study***
Setting 27	***A***	***1***	***Case study***

The baseline survey was sent to 152 participants across the 27 settings with a response of 54% (82 participants). The number of participants per setting ranged from two to fifteen and among those who responded to the baseline survey a little over half were early years practitioners (52%), just under a quarter were teachers (24%), a fifth were managers or deputy managers (20%), just over a tenth were teaching assistants (11%), and 7% were in a specialist role such as SEND. The post-test survey was sent to 153 participants but there was only a 29% response (45 participants) despite four chasing emails to individual participants from the evaluation and delivery teams and a reminder at the consolidation event. The vast majority of the participant sample was the same between baseline and post-survey points, and the sample increased slightly over this time because of changes in staffing at some settings, for example, new, replacement, or support roles. A total of 38 respondents answered both surveys so 46% of those who answered the baseline survey could be tracked between the start and end of the intervention.

Table 3 summarises how the data collection methods described above were used to address each of the research questions, along with suggestions for indicators as to how they have been assessed.

Table 3: Research question and data collection method summary

Research question	Indicator	Data collection method
Evidence of promise		
Is there evidence of an improvement in practitioners' practice and behaviour to better support children's early development?	***Self-reported change in practice by practitioners*** ***Setting managers' perception of change in practice***	***Surveys*** ***Case study visits and telephone interviews***
Is there evidence of an improvement in practitioners' understanding of pupils' competencies and skills?	***Change in understanding, based on practitioner self-report*** ***Setting managers' perception of change in understanding of practitioners***	***Surveys*** ***Case study visits and telephone interviews***
Do participants perceive a change in work practices from engaging in the intervention?	***Percentage of practitioners perceiving change in work practices*** ***Change in work practices based on practitioner self- report***	***Surveys*** ***Case study visits and telephone interviews***
Do participants believe that the intervention has had (or will have) benefits to children compared to previous practices?	***Percentage of practitioners who agree intervention has benefits for children*** ***Benefits to children perceived by practitioners and setting managers***	***Surveys*** ***Case study visits and telephone interviews***

Are there any unintended or negative effects of the intervention?	***Type and prevalence of any negative or unintended effects***	***Case study visits and telephone interviews***
Are practitioners more confident that they are correctly targeting children in need of support?	***Change in percentage of practitioners who agree they are confident in correctly identifying children in need of support*** ***Perception of change in confidence in practitioners and settings managers***	***Surveys*** ***Case study visits and telephone interviews***
Feasibility		
Is the intervention feasible for practitioners to engage in and implement?	***Experiences of practitioners and setting managers in relation to the intervention*** ***Issues covered in providing support*** ***Frequency of use of apps***	***Case study visits and telephone interviews*** ***Observations of support visits and AfC summaries of the visits*** ***Analysis of Toolbox data***
Can the intervention be delivered within routine practice?	***Participants' views on ease of using app within routine practice*** ***Frequency and timing of use of apps***	***Surveys*** ***Case study visits and telephone interviews*** ***Analysis of Toolbox data***
What are the facilitators/barriers to the programme?	***Participants' reports of facilitators and barriers***	***Case study visits and telephone interviews***
What level of support is needed to enable practitioners to effectively interpret assessment data?	***Type and extent of support required by practitioners*** ***Participant satisfaction with support received***	***Case study visits and telephone interviews*** ***Observations of support visits*** ***Surveys***
What level of support is needed to enable practitioners to effectively change their practice based on this data?	***Type and extent of support required by practitioners***	***Case study visits and telephone interviews*** ***Observations of support visits and analysis of delivery team summary of support visits*** ***Surveys***
What is the optimal / most feasible frequency and timing of the assessments within the school term?	***Practitioners' reports on experiences regarding timing/frequency of use*** ***Frequency and timing of use of apps***	***Case study visits and telephone interviews*** ***Analysis of Toolbox data***

Scalability		
Do the training and resources constitute a standard intervention that can be replicated and scaled up?	***Extent to which training could be standardised and scaled up*** ***Participants' experiences of training and using the app***	***Observations of training, consolidation event, and support visits*** ***Review of materials*** ***Feedback forms from training days/ consolidation event*** ***Case study visits and telephone interviews***
What, if anything, should be updated or changed for a trial evaluation (for example, theory, materials, procedures)?	***Participants' experiences of the programme***	***Case study visits and telephone interviews*** ***Surveys***
What will the cost of the intervention be to settings (including direct costs of training and indirect costs, for example, of interventions bought as a result of the assessments and costs in terms of staff time)?	***Direct costs of provision of training and resources as reported by delivery team*** ***Additional costs reported by settings (direct, indirect, and time required)***	***Cost information collected from delivery team*** ***Case study visits and telephone interviews***

Qualitative, survey, and child app data was used to triangulate learning. The data was analysed as follows.

Qualitative data analysis

Interviews were digitally recorded with the agreement of participants and transcribed verbatim. We analysed the data using a 'framework' approach, drawing themes and messages from an analysis of interview transcripts, observations of training and quarterly monitoring sessions, and other materials collected by evaluation and project teams. Framework is an Excel-based qualitative analysis tool that ensures that the analytical process and interpretations from it are grounded in the data and tailored to the research questions. Framework allows full within-case analysis (looking in detail at each individual case) and between-case analysis (comparing individual cases and groups of cases).

Analysis of survey data

Analysis of the practitioner surveys was intended to be light touch looking at change over time for practitioners in the key outcome areas and identifying examples of existing practice or assessment tools. As only 38 respondents answered the surveys at both time points, any quantitative analysis has been descriptive using SPSS and no statistical tests have been used.

Analysis of child app data

Child app data, collected through the EYT app and analysed in Stata, also provided a potential means of monitoring children's progress over the course of the intervention. Given the stage of the evaluation, the main purpose of this analysis was to explore the frequency and use of the apps rather than to assess evidence of promise. We consider use and frequency of the three apps in which practitioners were trained. We also looked for any variation across settings. The analysis is based on app data covering the entire period of the intervention; data was provided by the delivery team at the end of the intervention (August 2019), with an interim data file also provided midway through the study (April 2019) for preliminary analysis. Further details on the analysis of the app data are provided in Appendix 5.

Timeline

The timeline for the pilot is outlined in the table below.

Table 4: Timeline

Date	Activity	Responsibility
Aug 2018	***Set-up meeting***	***Delivery team and evaluation team***
Aug–Sep 2018	***Recruitment of settings including preparation and signature of MoU and other recruitment materials***	***Delivery team and evaluation team***
Sep–Oct 2018	***Usual practice questions and baseline setting survey***	***Evaluation team***
Oct 2018	***Delivery of training for practitioners*** ***Observations of initial training sessions***	***Delivery team*** ***Evaluation team***
Nov 2018	***IDEA workshop 1—feedback on training***	***Delivery team and evaluation team***
Oct–Dec 2018	***Delivery of follow-up training*** ***Observations of follow-up training***	***Delivery team*** ***Evaluation team***
Oct 18–June 2019	***Programme delivery***	***Delivery team***
Jan–Feb 2019	***Stakeholder feedback on ToC model***	***Evaluation team***
Feb–Mar 2019	***Wave 1 case study visits and telephone interviews***	***Evaluation team***
Mar 2019	***IDEA workshop 2—stakeholder feedback integrated***	***Delivery team and evaluation team***
Jan –Mar 2019	***Delivery of first support visits*** ***Observations of first support visits***	***Delivery team*** ***Evaluation team***
April 2019	***Collection of EYT data point 1***	***Evaluation team***
April–June 2019	***Delivery of second support visits*** ***Observations of second support visits***	***Delivery team*** ***Evaluation team***
May–July 2019	***Wave 2 case study visits and telephone interviews with other settings***	***Evaluation team***
June 2019	***Consolidation event***	***Delivery team and evaluation team (observing)***
June–July 2019	***Post-tests of practitioner survey***	***Evaluation team***
July 2019	***Delivery of final support visits*** ***Observations of final support visits***	***Delivery team*** ***Evaluation team***
Aug 2019	***IDEA workshop 3—final session with review of year and cost data collected***	***Delivery team and evaluation team***
Aug 2019	***Collection of EYT data point 2***	***Evaluation team***
July–Aug 2019	***Analysis of project and evaluation data***	***Evaluation team***
Sep 2019	***Presentation of findings***	***Evaluation team***
Sep- Nov 2019	***Report writing***	***Evaluation team***
Nov 29th 2019	***First draft of evaluation report***	***Evaluation team***

Findings

Participants

In total, 27 settings took part in the pilot study. Table 5 provides an overview of the characteristics of these settings. They included some Action for Children nurseries (six), but also other providers. Almost half of the participating settings were private, voluntary and independent (PVI) providers, with a further ten primary schools and four maintained nursery schools taking part.

Delivery was organised around three main regions broadly covering the South (covering Barnet, Hammersmith, Fulham, Essex, Newham, and Bexley), the North-East (covering North Hull, Gateshead, Northumberland, and Stockton) and the Midlands (Birmingham, Derby, Telford, and Wrekin). Most of the settings were located in urban areas, with 14 located in 'urban city and town' areas and seven in an 'urban major conurbation'. The remaining seven settings were located in rural areas.

Around half (14) of the settings had been rated as 'good' at their most recent Ofsted inspection,[3] with a further four rated as 'outstanding' and two as 'satisfactory' or 'requires improvement'. No Ofsted inspection report was available for seven of the settings.[4]

Settings tended to be located in more deprived areas (which is as expected as they were the areas targeted for recruitment), although the sample included settings from both more and less deprived areas. Table 5 shows the number of settings in each decile of the Index of Multiple Deprivation (IMD). On this basis, 13 of the settings were located in the two most deprived deciles, while two settings were located in the two least deprived deciles.

Table 5: Characteristics of participating settings

	Number of settings	Per cent
Ofsted rating[1]		
Outstanding	4	15
Good	14	52
Satisfactory	2	7
Missing	7	26
Setting type		
Primary school	10	37
Maintained nursery school	4	15
PVI provider	13	48
Region (government office region) [2]		
East	3	11
London	4	15
North East	8	30
West Midlands	9	33
Yorkshire and the Humber	3	11
Urban-rural classification[2]		
Urban city and town	14	52
Urban major conurbation	7	26
Rural town and fringe	3	11

[3] For primary schools, where available we use the rating for early years provision within the school. For nursery schools and PVI providers, we use the rating for overall effectiveness.
[4] This may apply to more-recently established providers that have yet to be inspected, those awaiting the publication of a report after a recent inspection, schools that have recently become an academy and are awaiting inspection, or 'outstanding' schools that have become an academy but may be exempt from routine inspection. It is also the case for those in a childminding agency that have a different inspection system.

Rural village	2	7
Rural hamlet and isolated dwellings	1	4
IMD decile[2]		
1	9	33
2	4	15
3	5	19
4	1	4
5	0	0
6	2	7
7	0	0
8	4	15
9	1	4
10	1	4

Notes and sources:

1. Ofsted ratings for primary schools and nursery schools: State-funded schools and inspection outcomes as at 31 August 2018; for PVI providers: childcare provider-level data as at 31 August 2018.
2. Identified on the basis of setting postcodes using the ONS Postcode Directory, February 2019.

No settings withdrew from the evaluation. However, some settings did not use one or more of the apps in all three terms, including one setting which did not use any of the apps in the third term, as discussed in more detail later in this report (see section on Implementation and Delivery).

Evidence to support theory of change

This section seeks to establish whether the piloted intervention showed evidence of promise, and to what extent it can find evidence on the theory of change as outlined above.

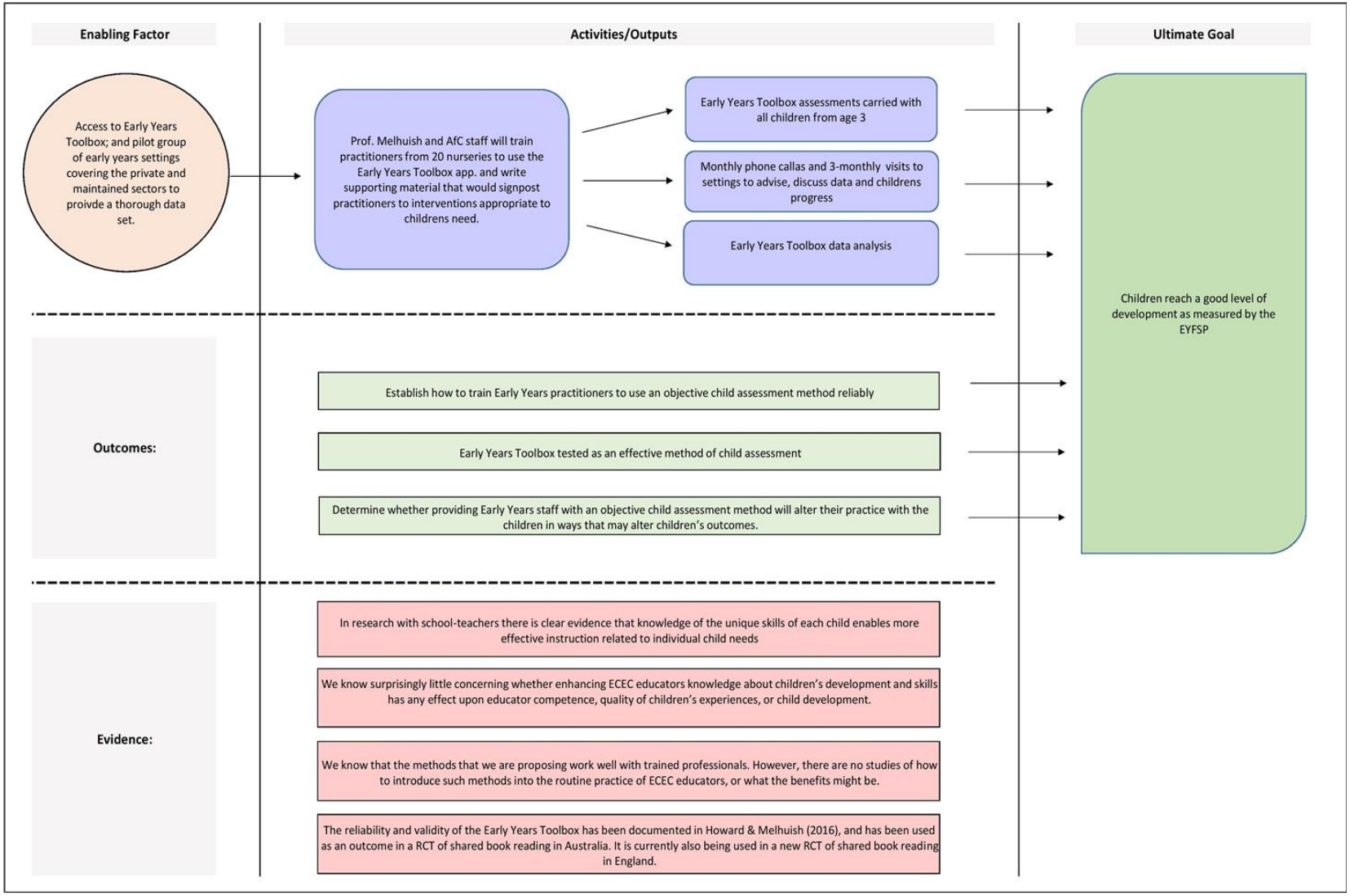
The section is structured as follows:

- first, we outline the theory of change of the intervention;
- second, we assess whether practitioners changed their practices and behaviour to better support children's early development;
- third, we explore whether practitioners believe the intervention had (or will have) benefits to children, compared to previous practices; and
- fourth, we explore whether there are any unintended or negative effects of the intervention.

Theory of change

Prior to the evaluation, Professor Melhuish and AfC had developed an initial theory of change outlining how the intervention was expected to impact on pupil outcomes as measured by the Early Years Foundation Stage Profile (EYFSP). Figure A presents this: here, intervention activities (blue boxes) and expected outcomes (light green boxes) are predicted to effect positive changes in children's development levels (bright green box). As noted above, the evaluation and delivery teams then further developed this model through a series of three IDEA workshops across the pilot and evaluation periods. The final two workshops used Connell and Klem's (2000, cited in Humphrey et al., 2016) criteria for assessment of quality of theory of change. Each IDEA workshop provided an opportunity to share updates on delivery of the intervention, feedback from the ongoing evaluation, and to review the theory of change. Through this iterative process, the theory of change was able to evolve alongside the development of the intervention.

Figure A: Initial theory of change model



The first IDEA workshop took place in October 2018 within a month of the initial practitioner training days for the Early Years Toolbox but before the follow-up training visits to settings to check how the training was being cascaded to other staff and practitioners who had not attended the training, and to provide support with this. This first workshop included discussions regarding feedback from practitioners on the initial training sessions as collected by AfC, findings from the baseline survey of practitioners, and a review of the programme materials. The theory of change was then assessed to identify any areas where it could be revised to better reflect the aims, processes, and assumptions underlying the intervention and any changes to delivery since the start of the pilot.

The second IDEA workshop occurred in March 2019, five months after the first workshop, and included discussion of anonymous expert stakeholder feedback on a revised draft of the theory of change model. The experts were specialists in child development and literacy drawn from the fields of education and psychology and had been contacted in January 2019 and asked to feedback on the latest version of the theory of change model. They were provided with a research briefing about the Early Years Toolbox, a feedback form, and the theory of change model. The feedback was received in February and collated by the evaluation team into anonymous feedback around key themes and issues.

Overall, the expert stakeholders were quite positive about the pilot and agreed that there was a need to explore how enhancing early years practitioners' knowledge of children's development and skills could inform and possibly benefit their practice. They also felt that the timelines and outcomes proposed were realistic. However, the experts highlighted that more detail was needed in some areas of the theory of change model to clearly map out how the intervention was expected to effect change.

- The experts highlighted that they felt that the improvements in child outcomes predicted by the model would only be possible if staff behaviour changed as a consequence of the intervention. They requested more detail on expectations regarding changes in staff behaviour and how staff would use the assessment data from the Toolbox apps. The experts felt that, in order to achieve this, additional resources and support would be needed to help staff identify appropriate actions after using the apps.
- As well as wanting to see more detail on the specific behaviour changes expected among staff using the apps, the experts requested clarification on what areas of children's development would be predicted to improve as a result of this behaviour change and how these outcomes would be measured. For example, whether this would reflect the areas assessed by the individual Toolbox apps—vocabulary, mathematics, self-regulation, and social development. Similarly, staff confidence and accuracy in assessing children could be a relevant outcome for inclusion in the model.
- The experts suggested including more detail on the intervention activities in the model and specifying how consultants would be trained to implement this with fidelity as well as how quality assurance for the subsequent cascading of training within settings would work. Aspects of this included possibly quantifying the number and frequency of visits from trainers, and stating specifications regarding which interventions/resources were identified as useful for staff to use with children after assessment, for example, evidence-based approaches.
- The experts thought that the model would benefit from a separate section on inputs required for the activities which could specify what resources settings would need to implement the intervention as intended such as staff time required to attend training or deliver assessments.

The third IDEA workshop took place in August 2019 within a month of the delivery of the pilot being completed and five months after the second workshop. The delivery team shared updates regarding the conclusion of delivery of the intervention and feedback around costs and resourcing, and the evaluation team shared initial feedback from the post-intervention survey, practitioner interviews, and case studies. The delivery team then shared the theory of change model which had been revised to incorporate feedback from the expert stakeholder panel as discussed at the previous workshop, and the teams discussed any further changes needed with regards to final delivery and understanding of the intervention.

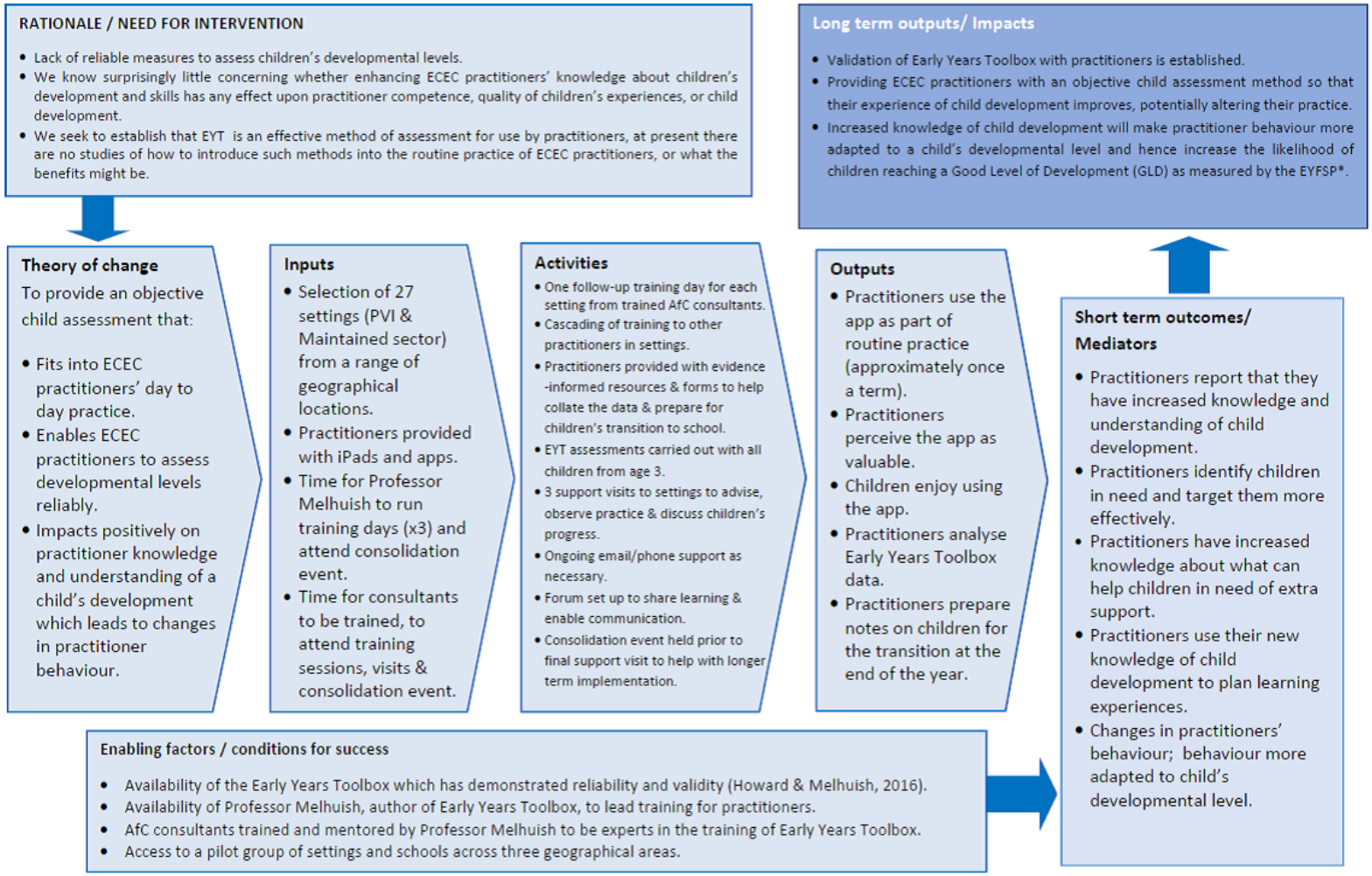
The final theory of change model (see Figure B) incorporated further changes made after the feedback from expert stakeholders and discussion at the final IDEA meeting as well as changes in the EEF IPE guidance. Key changes between the initial and final theory of change model included:

1. Overall, the theory of change model became more specific with regards to what the intervention intended to achieve and the causal mechanisms that would potentially lead to the outcomes specified.
2. The evidence section in the initial model was replaced with a section entitled 'Rationale/Need for Intervention' to outline the purpose of what the intervention was intending to achieve.
3. An 'Inputs' section was created in the final model that included elements such as time needed to deliver/participate in the training days and iPads.
4. Changes were made to the 'Activities' section to provide further detail and reflect changes to delivery as the pilot developed. Examples include: inclusion of an online forum, which was added after discussions in the first IDEA

workshop; a follow-up training session with AfC consultants was added in winter 2018; a consolidation event in June 2019 was added so that settings could share what they had learned while participating in the pilot; and more detail was added on the supporting materials for staff.

5. A section was added to specify outputs from the intervention such as whether or not children enjoyed using the apps.
6. The outcomes section was revised to focus on outcomes from the intervention overall rather than intermediate outcomes regarding delivering the intervention as part of a pilot.
7. A section detailing 'Conditions for Success/Enabling Factors' was added. This included elements such as Professor Melhuish's continued involvement in the training of consultants and staff for future roll-out of the intervention.
8. The mechanisms for change have been outlined in a section on 'Short-Term Uutcomes/Mediators'.
9. Long-term outcomes or impacts have been summarised in a separate section including changes to staff behaviour that are expected to increase the likelihood of children reaching a 'good level of development' as specified by the EYFSP.

Figure B: Final theory of change model



*Good Level of Development (GLD): "Children are defined as having reached a good level of development at the end of the EYFS in the reception year if they have achieved at least the expected level for the ELGs in: the prime areas of learning - personal, social and emotional development; physical development; and communication and language; and the specific areas of mathematics and literacy." (https://www.gov.uk/government/publications/early-years-foundation-stage-profile-handbook)

Evidence of change in working practices

This section seeks to establish whether the piloted intervention showed evidence of promise and to what extent it can find evidence on the theory of change as outlined above. It explores the following research questions regarding changes in working practices:

- Is there evidence of an improvement in practitioners' understanding of pupils' competencies and skills?
- Are practitioners more confident that they are correctly targeting children in need of support?
- Is there evidence of an improvement in practitioners' practice and behaviour to better support children's early development?
- Do participants perceive a change in work practices from engaging in the intervention?

Improved understanding of pupils' competencies and skills

In general, the interviews with practitioners and senior staff within settings identified limited change in their understanding of pupils' skills and competencies with some reporting more substantial impacts. There were a few settings that noted no impact, which tended to be ones where the practitioners felt satisfied with their existing methods of assessment, and these settings were often schools. Practitioners in these settings typically felt that they already knew their children well and that the Toolbox did not enhance this. Several of the settings were Action for Children nurseries, but there was a range of views and experiences among these settings regarding the Toolbox. Perceptions of impact and/or learning from the Toolbox intervention appeared to be more influenced

A few settings that noted no improvement in this area said that they still found it reassuring to find that results from the Toolbox matched with their understanding of the children's abilities gained from their existing assessments, such as observations. These practitioners who noted no impact tended to highlight that it was a useful piece of objective information which supported their observational data and this had fed into regular assessment reports and pupil profiles.

> *'We know where the children are, because we assess quite regularly but it was nice to have that extra just to back up what we were saying, with our evidence.'* Early Years Coordinator, Setting 18, Region B

> *'I think it's made them [practitioners] feel that what they're doing already is working really well, so it's just given them that little bit of reassurance.'* Nursery teacher, Setting 16, Region C

Among those settings interviewed that had noticed an improvement in this area, the majority were cases where the assessments had revealed gaps in the children's knowledge. However, there were also some cases where the apps had revealed individual children to have better knowledge of maths or English than the practitioners expected. This aspect of the Toolbox was seen as particularly valuable in settings where practitioners are less experienced, according to senior managers, in spotting gaps in the children's knowledge.

> *'There was one child who started in September, so we hadn't got to know him really well, who it turned out that had quite a lot of mathematical knowledge that we didn't necessarily know.'* Nursery teacher, Setting 16, Region A

> *'The child was very quiet and we didn't realise how good his vocabulary was, so where practitioners aren't so great at assessing all children, I think it can plug gaps.'* Practitioner, Setting 14, Region A

Practitioners in the interviews were asked to consider the impact of the different apps on their knowledge. Settings that had noticed an impact in their understanding of the children's abilities were more likely to report this about the numeracy app compared to the other two apps. Many highlighted it was useful that assessment scores were presented in sub-categories of maths (such as ordinality, cardinality, and patterning) rather than just one overall score as in the vocabulary app. Practitioners welcomed this as it helped to identify knowledge gaps for individual children in the different sub-areas. Some practitioners noted that some of the categories were not covered in Development Matters[5] such as cardinality in mathematics, so this app was a useful addition to existing practices. Maths was also a topic that many settings, especially in the private and voluntary sector, felt was not strong in their practice and they welcomed the opportunity to improve in this area.

The vocabulary app was less well-received by practitioners in the interviews than the numeracy app. Those that liked the app commented that it was a simple assessment tool, simpler than the numeracy app, and that it gave them a quick idea of where the children were in terms of language and vocabulary knowledge. Some practitioners also said that it was valuable one-to-one time, which made practitioners aware not just of gaps in vocabulary, but also certain pronunciation issues. However, some practitioners in the interviews said it felt like a 'crude measure for vocabulary',

[5] Development Matters is a free, downloadable and non-statutory guidance by the British Association of Early Childhood Education to support practitioners in implementing the Early Years Foundation Stage (2012).

and questioned whether it can capture the complexity of language development including children's ability and progress. This view was expressed particularly among settings that had other, more detailed language assessment tools such as WellComm.[6] Many practitioners also felt that the vocabulary app was less able to pick up differences in the different learning sub-areas, which was the aspect of the numeracy app they had found the most useful.

Settings, in the interviews, had mixed views about how much the self-regulation app added to their understanding of the children. As with the other two assessment methods, settings where practitioners expressed confidence in their existing knowledge of the children felt that completing the questionnaire did not add to their understanding. These settings were mostly schools. Sometimes this was because they already worked with self-regulation in the setting or used similar type of assessment tools, including self-esteem indicators that were considered to be better and more widely recognised when looking at referrals.

In other cases, practitioners made the point that it did not tell them anything new because the information they inputted was based on their subjective observations.

> *'I found that less useful, just because you need to know the children to be able to answer it. So, if you could answer it, you sort of know what sort of a learner they are, where their strengths are … I could have told you that, because I put the answers in.'* Nursery teacher, Setting 13, Region B

> *'The only concern was sometimes I felt it told me what I already knew. So, he is low on behavioural, but high on social skills, for instance. I would be like, "I know that of that child."'* Nursery teacher, Setting 25, Region A

Other settings in the interviews were more positive, and in contrast to the respondents above, they implicitly argued that while they may have possessed the knowledge about the children, completing the assessment had given them a chance to reflect on children's self-regulation. This included cognitive, behavioural, and social self-regulation, as well as social behaviours. This was described as good 'food for thought' and generally made them more aware of self-regulation, and made them reflect on how to self-regulate.

> *'That's working really well, because it makes you sit down and think.'* Pre-school room leader, Setting 21, Region B

The settings where the self-regulation app seemed to be particularly valued were those where practitioners had completed the assessments in pairs or groups of practitioners. Most of these settings had done it in pairs, for instance two practitioners, or a practitioner with a manager or room leader. In one setting, they had discussed about ten children as a whole team on an INSET day. These settings said the discussions were useful and made them think about the ability and progress of individual children, sometimes in more depth than they would usually do. Some of these practitioners also highlighted that the questionnaire covered many of the same things as the Early Years Foundation Stage profile, which meant that it complemented other reports.

> *'[It was] very thought-provoking for ourselves because obviously we have to do it, so we're just thinking about the child and really getting to the bottom.*' Nursery manager, Setting 2, Region A

> *'I do think if you've done it with more than one person it's probably more accurate than if you just do it yourself.'* Practitioner, Setting 24, Region A

Some also said that the self-regulation app reassures practitioners that they know their children, which feeds into their reports as formal assessment data. A practitioner said:

> *'It's okay ... it lets us know how well we actually know our children, which is quite nice for us.'* Practitioner, Setting 12, Region C

Improved targeting of resources to individual children

In the interviews, there were some settings where the Toolbox data had been used to introduce activities that target individual children based on their needs. In some of these settings where the child was found to have abilities that were below the average for their age, practitioners would plan additional interventions to support the child in the areas that required improvement. In particular, this was found to be the case with the numeracy app that broke down the score to the different learning sub-areas and helped the practitioners to plan specific activities focusing on areas where the child had struggled.

There were also a handful of cases where the setting had identified a speech and language support need in a child when using the vocabulary app. In these cases, practitioners felt that the Toolbox had been beneficial because having

[6] WellComm is a toolkit designed by GL Assessment that enables practitioners to identify early years children in need of speech and language support.

a numerical assessment score of the child's ability made their case for extra support stronger. In some cases, the assessment had revealed a support need earlier than might have been the case without the assessment.

> *'There was a little girl that we did it the first time and we did it a second time and she didn't make any progress whatsoever, we knew she's behind and she's not making much progress. We didn't realise she hadn't made any progress in them particular areas.'* Nursery teacher, Setting 18, Region B

On the whole, however, the evidence of improved targeting of resources was limited in the interviews. Practitioners did not explicitly discuss barriers to making more improvements in this area, but the qualitative interviews indicated certain barriers to implementation that may explain why this was the case. For one, many practitioners felt that a key aspect missing from the Toolbox was that it does not offer advice on what to do with the results once you have completed the assessment. One practitioner compared the app to WellComm, which provided ideas for next steps based on the assessment results including teaching and activity resources. Another practitioner also made this point and said this was partly because the score was not broken down into different vocabulary concepts:

> *'At the training, I thought "great!", but now I'm not clear about how the data is supposed to be used, compared to the numeracy app. Especially because you don't have a breakdown of different concepts like the numeracy app, so you don't know what to focus on.*' Nursery manager, Setting 2, Region A

Secondly, only some settings in the interviews used the data to target individual children. In many settings, practitioners considered the results for the whole cohort and used these to introduce group-based activities. Where settings did use the data for individual children, in many cases settings used it to target 'outliers' (children who over- or under-performed in the assessment) as opposed to each individual child. It is worth noting, however, that practitioners did not view this type of targeting of resources for outliers (as opposed to using it with all the children individually) as negative or unproductive.

Practitioners were also asked about their self-reported ability *to identify children who are most in need* in the surveys. Practitioners reported that they had more confidence in self-regulation between baseline (82%) and post-treatment (92%) surveys, alongside social development (87% to 92%). There were no increases in practitioners' confidence to identify children most in need of support in language and numeracy skills. However, self-reported confidence levels were already high in the baseline survey so there was limited range for improvement, and the number of respondents who answered the questions in both surveys was quite small (37–38) so these findings should be interpreted with caution.

In addition to this, practitioners were asked to rate their self-reported confidence *in assessing children*. The post-treatment survey showed increases in practitioners' confidence to formally assess children's social development (84% to 100%, n = 38 each) and self-regulation (79% to 90%, n = 38 each) as well as practitioners reporting a large reduction in the need for additional help in assessing all areas (a 34–38% reduction across the areas). Practitioners who agreed that they would like more help with assessing children's development reduced from over half of respondents at baseline (58–62%, n = 37–38) to around a quarter of respondents at post-treatment (23–29%, n = 34–35) across the areas assessed by the Toolbox.

Improved practice and behaviour

Among practitioners surveyed after delivery of the Toolbox programme, a majority felt that the Toolbox had helped improve their practice (61%, n = 44)—most commonly by using the results to inform their planning (64%, n = 45).

In the interviews, a number of settings described how the increased knowledge of the children's abilities was used to inform and coordinate activities.

> *'You can see where the children are, and then you can differentiate accordingly with your activities.'* Room leader, Setting 20, Region B

> *'I think being aware of that [children's abilities], I think that's a huge impact really, and I think it's one that we'll really bear in mind when we're planning for the teacher-directed groups, and also during free flow when we're interacting with the children, we'll try and get those words in a little bit clearer.'* Nursery teacher, Setting 20, Region B

Some practitioners reported becoming more conscious of the needs of their children following the intervention.

> *'It's just made me more aware of my children and where they are. So I think it will help people, I know we all do a brilliant job and we all do our focus time, and our next steps and our observations. But just those little things that you wouldn't normally think of doing, it just helps you to understand that they are not understanding certain things. So it makes me change the way I am speaking and teaching them.'* Supervisor, Setting 22, Region B

Looking at findings for individual apps, there were several examples of practitioners improving their practice and behaviour on the basis of the numeracy app assessment. Settings in the interviews appreciated how the app, compared to other assessment methods they use, highlighted specific areas that required improvement, giving settings 'more pinpointed areas to look at'. For some settings, seeing these categories had prompted them to focus on improving teaching practices in specific areas, for example on positional language, patterns, and sequencing.

> *'A few of ours struggled with sequencing so that is now in our maths planning time. We've put into place activities to help children with sequencing. Even lining up to go outside, we slot it in where we can.'* Practitioner, Setting 9, Region B

Many settings described how, after seeing results from the numeracy app, they have become more mindful to incorporate maths language and activities to their daily practice and think about the different areas of maths again.

> *'We use it everywhere now, rather than only in certain places. We've added things to areas, like tape measures and things like that, and they're there to be used now, they are actually used on a daily basis now.'* Nursery manager, Setting 19, Region B

Practitioners at some settings felt that data from the numeracy app and discussions with consultants about maths during their support visits had transformed their settings' approach to maths teaching. For these settings, the support visits played a crucial role in providing ideas through discussions during support visits by providing additional materials and suggestions for activities and, in some cases, by delivering bespoke training sessions during staff meetings. One setting described the change in the following way:

> *'We changed everything; some of my children have moved onto the addition and subtraction now. I would do it about once a week beforehand, but now I try and move maths into every activity we are doing, I will bring it in with sizes and positional language.'* Supervisor, Setting 22, Region B

The interviews showed fewer cases of improved practice as a result of the vocabulary app. Settings often felt it had a limited impact on their knowledge of the children, which they felt was already at a high level or covered by other assessment methods. Where it had made an impact on practice, it was used to inform planning. One setting introduced a new activity, 'vocabulary bag', first to the whole class but now also as an intervention for those children who had low scores in the assessment. Others described becoming more mindful to use the vocabulary from the app to reinforce the learning. One nursery teacher described the change in the following way:

> *'You see them not recognise something like a shell. That will then lead onto "I need to maybe do something more about shells". So it will highlight things to you about their learning as well, lead the planning.'* Nursery teacher, Setting 25, Region A

As with the vocabulary app, those practitioners in the interviews that felt they already knew their children well said the self-regulation had not led them to change their practice or behaviour. A handful of settings had introduced new activities based on the app. In one case, the self-regulation app had made the practitioner more aware of why a certain child behaved in a certain way and she subsequently put interventions in place. In another setting, a child was struggling with socialising and the practitioners created a 'thinking tent', which is a space in the classroom where the child went if they were overwhelmed by social situations.

As well as informing practice, there was also some evidence that using the Toolbox apps informed practitioners' knowledge or awareness of levels of knowledge among staff regarding specific topics. The majority of practitioners who responded to the post-intervention survey (64%, n = 45) indicated that they had used information from the Early Years Toolbox to identify areas where they would like to increase their knowledge. In the interviews, some practitioners reported that using the Toolbox had helped them to identify gaps in knowledge among staff at their setting or topics that hadn't been covered. One practitioner reported that using the numeracy app had helped staff at their setting to become more aware of different mathematical concepts and areas. Another practitioner reported that a colleague had needed some support to understand and respond to the questions in the self-regulation assessment app and they felt that other colleagues might also need support with this app.

During interviews, one nursery manager identified a further benefit to staff of using the Toolbox. She felt that the Toolbox was useful because it made her staff feel that the work they do is important. The manager used the grids of pupil data against norms for their ages to show her practitioners how the children had progressed, and this gave them a confidence boost.

> *'Just because a lot of people see us, "Oh, they're just babysitting, they're just playing in sand, playing in water", they don't see, they don't kind of understand how important it is. So it's handy for me to be able to say to the staff, "Look, what you're doing is working, they are improving, they are progressing—you're doing a great job, really."'* Nursery manager, Setting 19, Region B

Limitations to changes in work practices

The majority of settings in the interviews found the results of the Toolbox assessment helpful to their practice, at least to some extent. In the second wave of the evaluation there were slightly more cases of settings not finding the results useful, or only finding limited use for the results, although this was still a small number of settings. Reasons given for this included:

- practitioners feeling that the app data did not tell them anything that they did not already know;
- that there was insufficient guidance on what actions to take after identifying knowledge gaps; and
- concern that the app data may not be reliable due to some children growing bored or tired meaning that their scores would not be a fair reflection of their ability.

These issues are explored in further detail in this section. None of the settings described a negative change to their work practices—at most, it had made no impact. This is because practitioners tended to view the apps as quick and easy to use, and the children enjoyed them.

Practitioner-reported changes to work practices tended to relate to specific activities rather than overall changes indicating limited impact on changing practice at organisational level. In the survey, just under half of practitioners felt that the Toolbox had helped them to improve practice for all staff working with this age group at their setting (48%).[7]

The settings that took part in the interviews that reported no change in work practices were all attached to schools. Nursery teachers in these settings felt that the assessment results were not surprising and had not added to their understanding of the children's knowledge, therefore making little impact on their practice. These practitioners, and their senior leaders, tended to hold the view that as experienced early years practitioners they are able to spot gaps in the children's knowledge as it is. Some nursery teachers felt that the Toolbox might be more helpful in settings where practitioners are less knowledgeable about children's developmental stages.

Schools commonly felt that the Toolbox in its current format is missing follow-up guidance and resources on what to do with the assessment results.

> *'The main thing that was missing was the advice on what we could do about it, you know—the next step. Whereas our WellComm assessments, they come with a book of ideas that you can … when you can find the gap in the child's knowledge then you can look it up and it offers advice, and activities and things that you can do to fill that gap. Whereas obviously the app just gives you the score on whether they are below age or as you expected.'* Nursery teacher, Setting 15, Region B

It should be noted that the delivery team did develop resources after delivery of the intervention had started, such as additional literature for practitioners to read, but this material was not widely read and used, according to the interviews. Some further materials were developed by the support consultants on an ad hoc basis. However, the support consultants did not have a consistent approach to developing and sharing tailored resources with their settings, which may have had an impact on practitioner views of the support offered. For example, there appeared to be variation by region as to whether staff reported receiving tailored resources as part of their consultant support. This lack of a consistent approach to follow-up resources and guidance may have had an impact on practitioner views.

Some settings in the interviews described a limited change in their work practices. Typical examples of this were cases where settings had introduced some further activities based on the results, but because the results largely confirmed what they already knew, practitioners felt their overall approach had not changed. In one setting, the nursery teacher reflected that the Toolbox has been helpful in reiterating to staff that you can use advanced vocabulary with the children. Other settings felt that while the Toolbox had been helpful in informing planning and highlighting gaps in the children's knowledge, ultimately it was no more useful than the other assessment methods they used.

> *'It hasn't really truly changed how we do things, we've always identified what the children need, obviously, and this has just been like another tool to use, to highlight the areas of learning that we need to concentrate on.'* Nursery teacher, Setting 17, Region B

Benefits to children

Both qualitative and quantitative approaches were used to explore benefits practitioners identified for the children as a result of using the Toolbox. In the survey, around half of practitioners felt that the Toolbox had positively impacted learning and/or behavioural outcomes for children who needed more support (55%) and that it had greater benefits for children than other approaches (54%).[8]

[7] Please note that a small number of respondents answered this question (44) so this finding should be interpreted with caution.
[8] Please note that the number of respondents who answered these questions were quite small (44 and 43 respectively) so these findings should be interpreted with caution.

In the interviews, settings expressed mixed views about how much of an impact the Toolbox has had so far on the children. Many noted limited or no impact, especially those that had not changed their practice to a significant degree, which were more likely to be schools than other types of settings. A couple of the settings were unable to cite positive impacts but noted that there have not been any negative impacts either, as the children enjoy using the apps. Others noted no significant impact apart from a single child who has benefitted from interventions. Some settings were withholding judgement about the impact on children before they saw the final Toolbox data.

Where settings, as reported in the interviews, had noticed a positive impact, there were varied examples of that. One setting reported that it had noticed that the children's concentration was a lot better while another one said that assessment results on the vocabulary app had shown consistent progress. A third setting said that they have set tasks during play time that they usually would not have done and since then children have been more inquisitive about positional language. Lastly, one setting that had significantly changed their approach to maths teaching described the impact on children in the following way:

> *'Their maths has improved immensely compared to others that we've had previously, I think because we're more confident now, with teaching it, I think they're a lot more confident to try it.'* Nursery manager, Setting 19, Region B

In most cases, settings in the interviews were unable to articulate the expected impact on children over the following 12 months. Two settings said that they expected the Toolbox to have a similar impact each year. Another setting said that if the Toolbox helps practitioners reduce their workload, it will allow them to put more time and effort into activities such as teaching and reading. They also anticipated that if they focused on the bottom 20% of children, the Toolbox would help inform planning for those children.

Negative or unintended consequences

Generally, settings, in the interviews, did not report many negative effects resulting from the Toolbox and where they did they were varied in nature. One setting felt that the assessment is setting up their EAL and high needs children to fail, and some practitioners felt that the assessment could be challenging for EAL students. However, one practitioner reported that their EAL students had performed better than they expected, which, she speculated, could be attributed to the kind of vocabulary that the EAL children were learning in English. A handful of settings noted the time it takes to complete the assessments, and the time out of the classroom, as changing the running of their planned sessions. A couple of settings said that some children, of both high and low ability, reacted negatively to getting the answers wrong.

One unintended effect of the Toolbox was settings using apps as teaching tools as well as assessment tools. In the case of the numeracy app, some practitioners found that the real potential of the app was as a teaching tool and expressed concerns that they were not allowed to provide teaching and instruction to pupils during the game. For instance, a setting manager said:

> *'It would be a prime opportunity, if they got it wrong, to then explain it to them, but we can't. We're missing opportunities to teach children by doing this application.'* Nursery manager, Setting 10, Region C

Another setting that had effectively dropped out of the programme in terms of using the apps, said it still valued the numeracy app as an educational tool, especially targeting at the higher-achieving pupils:

> *'We didn't end up using the maths app as assessment tool so much, but to offer the children the chance to have a go. It was another way of practising their maths, because they could interact with it … We used it more with some of the higher-ability children. It brought something a little bit extra to those children as you worked your way up with the maths. Once things got more interesting for them, it gave them a bit of a challenge.'* Nursery teacher, Setting 15, Region B

Some practitioners in the interviews also saw the vocabulary app as more valuable as a teaching tool, though this was not as common a view compared to the numeracy app. Those settings said it would potentially be similar to flashcards, and could be used as a teaching tool. One practitioner said it could be useful as it contained some 'unusual words' which children would not have come across before. At least one practitioner had used it as a teaching tool though she knew this was not the purpose of the app:

> *'I did say, rightly or wrongly, when they got an answer wrong in the vocabulary one, I would say "that was a good guess" and they got another chance, and I'd say, "The real word for it is a fountain." So they found that quite interesting because they were learning something from their mistakes whereas I feel the number one not so much.'* Practitioner, Setting 24, Region A

On a broader point, practitioners in the interviews expressed mixed views about whether or not the Toolbox assessments were a good use of their time. A few settings felt that it was time well spent, one-to-one with a child. There were also arguments made to the contrary. At one setting, practitioners said that they were not sure if this is the best

use of their time and that they prefer doing maths during play time because they find that more productive. Another setting questioned whether the type of one-to-one time the assessment offers is valuable:

> *'You are spending a lot of time doing it with them and I don't know how cost-effective that time is spent sitting one-to-one with the child doing these games, because you're not actually teaching them anything when you're doing them; it's pure assessment.'* Practitioner, Setting 1, Region A

Feasibility

This section assesses whether the programme is likely to be practical and feasible for providers to engage in and implement. In particular, this section describes withdrawal and participation rates among settings, and explores to what extent participants believed the programme requirements had been realistic and feasible to implement. In addition, barriers and facilitators to delivery experienced by settings and practitioners are described.

Specifically, the section is structured as follows:

- first, we describe practitioners and settings' views on the support provided as part of the intervention and assess whether this level of support enabled practitioners to deliver the intervention;
- second, we explore to what extent the intervention can be delivered within routine practice, including in relation to the expected frequency of assessments, number of children, time and space requirements, and other factors;
- third, we explore whether the numeracy and vocabulary games, as intended by the design, were perceived by pupils as a game and a fun activity rather than as an assessment;
- fourth, we explore whether practitioners felt confident and had the appropriate level of support in collecting the assessment results; and
- fifth, we identify barriers and facilitators to delivery experienced by the participating settings and practitioners.

Training and support

This section describes our findings on practitioners' views on the support provided as part of the intervention, including the initial training day, cascading of training day information, the follow-up training session, the support visits, consolidation event, and other support mechanisms. The section assesses whether this level of support made practitioners prepared to implement the Toolbox intervention.

Initial training day

Almost all of the settings sent two practitioners to the initial training day who were then responsible for cascading the information to relevant practitioners in their own setting. The evaluation team observed each initial training day in all three regions. During case study visits and phone interviews, we interviewed both practitioners who had attended the initial training day and were responsible for cascading the information as well as interviewing practitioners who did not attend the training day. Finally, the post-treatment survey included specific questions on the experiences of the initial training day.

Overall, all survey participants agreed that the training day was 'clear and useful' and that it had made them 'understand the aims of the intervention' and 'understand how to use the apps'. During the case study visits, practitioners said the initial training day was useful and interesting, and practitioners felt prepared to deliver the programme after the training. Two aspects of the training day were seen as particularly useful. First, the opportunity to try out the apps with children gave practitioners a better impression of how the intervention would work in practice and how pupils reacted to the games. This made practitioners more comfortable about implementing it in their own setting. It also gave them an opportunity to identify any immediate issues or concerns, which they could discuss with trainers face-to-face during the training day. It also showed practitioners and trainers some of the challenging aspects of the intervention, such as the tendency among practitioners give too much feedback and prompts during the games, which was typically discussed afterwards in the whole group, and the trainers highlighted the importance that practitioners put on the 'assessor hat' rather than the 'educator hat'.

Second, many practitioners said it was particularly useful to hear the perspectives of the trainer and co-designer of the app, Professor Melhuish, who explained the background and rationale for the intervention. Some said it was an 'eye-opener' how important early years' education and vocabulary is to later educational outcomes, while others said it reaffirmed what they already knew. Some said it made them reflect on the differences in exposure to language and vocabulary experienced by children from affluent and deprived backgrounds. Some also explained that it helped them understand why assessments were needed. Overall, many practitioners felt this was important for buy-in. For instance, one practitioner commented:

> *'I thought it was great, because he was giving us a reason to do it.'* Manager and practitioner, Setting 2, Region A

Most practitioners said all components of the training had been useful, and most offered no suggestions for how to improve the training day. The only main theme that was identified was that some practitioners said it would have been useful to have been provided with information at the initial training day about how to feed the assessments into planning and practice.

Other issues were identified as detailed below, but these were only mentioned by a few practitioners:

- Some practitioners commented on how the iPad training was conducted. A younger nursery manager in Region A said that trainers spent quite a long time going through the apps, though she considered them 'quite self-explanatory'. Similarly, a few practitioners in Region C described the training on the iPad as 'patronising' and noted that this was often the case for Early Years' CPD. Some of these practitioners, however, thought it may have been useful for older practitioners who may not be as confident with the iPad. This seemed to be confirmed in interviews with others who emphasised that the training day had made them confident about using the app.
- Some settings in Region A noted that the sheet with the grid for data collection had not been produced for their training day. This part of the intervention had not been clear at the time for practitioners from some of those settings. AfC subsequently produced the data collection sheet in time for the two remaining training days in the other regions, and it was introduced to Region A settings afterwards and during the follow-up training session.
- Some settings had problems with downloading the app ahead of the training day.
- Some settings in Region C said they had received mixed messages in their respective groups about whether to accept certain words, such as 'daisy' as well as 'flower', in the literacy app, which had caused confusion.
- Practitioners in a specific setting in Region B noted that the initial training day ideally would have been closer to their setting. This is an issue which would likely be solved in a larger intervention with a larger number of participating settings in the treatment group, and potentially smaller regional hubs.

Cascading the information to other practitioners

Following the initial training day, practitioners were then expected to implement the programme in their setting. In the first instance, for most settings the next step involved cascading the information to colleagues. AfC had not provided any specific guidance on how best to cascade the intervention to other practitioners, but most settings had used a similar approach in which the practitioners who had attended the training day sat down and explained to other practitioners what the apps were about, and then demonstrated the use of the apps, and observed while each practitioner tried it out.

All training day attendees who completed the post-treatment survey said it was 'very easy' (56%) or 'quite easy' to cascade (44%, n = 18), and almost everyone felt 'confident to train other staff' (96%, n = 26) to implement the intervention. Similarly, during case study visits cascading was described as 'fairly straightforward' due to the simplicity of the apps. Some practitioners noted this was a positive, compared to other assessment tools which would have required more time to disseminate to other practitioners:

> *'There's not a lot of training really. It's quite straightforward.'* Manager, Setting 8, Region C.

> *'I demonstrated the app and showed them, but it's so straightforward they went with it anyway, they could figure it out … Even my one that absolute hates IT, she managed to do it.'* Deputy manager, Setting 9, Region B.

> *'They made it so easy. It wasn't as if it was going to be difficult for me to come back and train the other staff members to use it. A lot of tracking tools are quite complex ... so it's difficult to actually relay the information back to them, whereas this was really simple, and they could do tests as well before they used it with the children.'* Manager and practitioner, Setting 19, Region B.

This was reinforced by practitioners who had not attended the initial training. They generally said that the introduction by their colleague had been good and that the app was easy to use. The main challenge, which was acknowledged both by training day attendees and most non-attendees, was to make sure that practitioners put on the 'assessor hat' rather than the 'educator hat'.

> *'What I do find difficult is to sit back and not help them when they're doing it, because I think as a practitioner, you're there to help the children.'* Nursery teacher, Setting 26, Region A.

Many practitioners seemed to have put a specific focus on this aspect when cascading the information to colleagues. Though, in a few cases, the instruction on the initial training day seemed to have been taken too far in the sense that practitioners thought they were not allowed to give any instructions about how the game worked. For instance, a nursery teacher in Region A said it was particularly difficult to avoid intervening during the numeracy app when children started answering before the voice stopped talking. This strict interpretation was also apparent in another setting, where the manager who had not attended the training day, but had been handed over the lead of the programme from a departing practitioner, said her predecessor had given them the following instructions:

> *'Well, you sit the child down and you don't talk to them, and if they need to hear it again, you press that, and don't tell them the answers, don't encourage them, don't put it in a different way, and then once they get so many wrong or can't answer them, then the test will finish.'* Manager, Setting 22, Region B.

Generally, while practitioners said it was easy to cascade the information, it should be noted that there was a marked difference between the amount of information and training that different practitioners received during this intervention. While training day attendees received fairly detailed instructions on how to play the games and the background on the importance of assessment, the practitioners who did not attend the training, in contrast, typically received fairly light-touch instructions, usually focused only on the practical aspects of how to play the games. This seemed to have an impact in some settings. While this would not necessarily be true for a larger sample size, it was worth noting that in the few cases where parts of the intervention had been misunderstood, this usually involved practitioners who had not been involved in the initial training day, and the follow-up training and support visits had been crucial in making sure that these practitioners implemented the intervention as intended. The evaluation also observed that non-attendees seemed less likely to have seen any of the additional resources and the slides from the training (although these were shared by AfC with training attendees), and they seemed less likely to know that Toolbox was not simply about collecting assessment data, but also supposed to inform planning and practice. Finally, for some settings, particularly smaller ones, only two practitioners played the games, so it had not been necessary to cascade the information from the initial training day.

Follow-up training and follow-up support visits

Settings were supported with three follow-up visits by a consultant, one visit per term. The evaluation team observed a total of 11 follow-up training and support visits across the three terms. In addition, as part of the case study visits, we asked practitioners about their experiences of the follow-up training and support visits.

Generally, the follow-up training visit had been used to make sure that the intervention was set up appropriately and had given practitioners the opportunity to discuss any issues or concerns. The consultant had discussed progress with the settings, observed practitioners complete assessments with children, and provided advice when practitioners needed this to adjust their use of the app. This seemed mostly to be related to making sure all practitioners were only using the game for assessment rather than as a teaching tool. Some settings, particularly those in Region A, that had not been introduced to the data collection sheets during the initial training day, said the first support visit had been crucial in clarifying how to collect data. Some practitioners also said the first visit was useful for solving any technical issues and generally providing reassurance that practitioners were implementing the Toolbox intervention as intended. In at least one setting in Region A, the first support visit seemed crucial in making sure that Toolbox was being implemented correctly as the practitioners had played the games regularly with the children until then.

Subsequent support visits were more tailored to the needs of the specific setting, and experiences were more varied. While this tailoring was intentional and proven to be effective in some cases, the evaluation also identified that the variations in support were not only caused by the specific needs of the settings, but also by the different styles of individual trainers and their understanding of the aims of the support visits. Some settings, particularly in Region B, highlighted that the consultant had provided ideas, resources, and activities on how to support children identified as lower-achieving in certain areas, and more broadly how to improve maths and language teaching practices in the setting. For instance, a supervisor in Region B described how the consultant worked with her to figure out why a child had seen a drop in assessment and then given her suggestions on how to change practices accordingly. Many of the settings in Region B, in discussions with the consultant, had used the assessment scores to identify specific focus areas for the whole setting. Subsequently, these settings had focused on improving their practices around this particular area, for instance on self-regulation or maths, and they had really valued the support during the termly visits. This had also happened for some settings in Region C, which had subsequently focused on using more maths language in their everyday practice.

Settings in Region B also highlighted the inspiring conversations which had made them reflect on teaching and practices. For instance, a practitioner in Region B had valued the opportunity to bounce ideas off an experienced practitioner, particularly as her setting was run by a voluntary committee and other staff were not from a childcare background. A typical comment was:

> *'We had really inspiring conversations together. I just wish we could have that all the time, because it's so good, because sharing those ideas, it helps you to move your own thinking forward, about ways to use it within your own setting.'* Teacher, Setting 20, Region B

In contrast, particularly in Region A, it seemed that practitioners had not received the same amount of support and forward-looking advice on how to use the assessment to inform planning and practice, and sometimes the visits had been more about giving feedback about the apps than about providing support or advice. This was also observed in the evaluator's observation visits in which support sessions seemed more like feedback sessions and the Toolbox was

portrayed as a pilot intervention where the priority was to gather information about what was going well and badly. One practitioner in Region A commented:

> *'More than anything it's probably been more useful to [my consultant and AfC] than it has for me doing the follow up visit.*' Nursery manager, Setting 23, Region A

In these settings, the value of the visits seemed mostly to be about reassurance that practitioners were implementing Toolbox correctly and that the issues and problems they had experienced were shared by other settings. A nursery teacher said:

> *'It's nice to [have a support visit] because we got together as nursery staff with [the consultant] and it was just nice to discuss what each setting was doing, and I think across the board, most people were having the same sort of teething problems.*' Nursery teacher, Setting 26, Region A.

Overall, from our interpretation of how people spoke about the different elements of the intervention, as well as from practitioners' own accounts, the support visits seemed to be less important in the minds of practitioners in Region A than practitioners in Region B and, to some extent, Region C. It should be noted that the consultant in Region A had close and already-established relations with a number of the participating settings and visited them regularly for other reasons. Practitioners in those settings may have received more ongoing and regular support on Toolbox, maybe sometimes when and as needed, which may not have been picked up in the termly observations or during interviews.

Overall, while the initial training days were very similar across regions, the training and support visits varied in terms of the type and amount of support given. The final visits included a focus on sustainability of use of the apps and how these may help with children's transition to school. Of course, the visits were intended to be tailored to the specific needs of the settings, but sometimes the variations were fundamental in terms of how much support was offered, particularly to what extent the consultant provided ideas and advice on how to use the assessments to inform planning and practices. Any subsequent rollout of Toolbox should make sure that all consultants are instructed consistently about the purpose and aims of the support visits to make sure that support is tailored but consistent, and that settings are given the same high-quality support.

Consolidation event

The evaluation team observed the consolidation event in London in June 2019. The delivery team handed out a feedback sheet about the training during the event and the responses were shared with the evaluation team. Similar to the evaluation team's observations on the day, it showed practitioners to be engaged and that they valued the opportunity to share experiences.

Other support mechanisms

AfC provided a folder of physical Toolbox resources including materials on supporting development in the three key areas covered by the three apps. Most practitioners did not seem to have used these materials, or did not recall whether they had used them in the beginning of the intervention; most could not remember the content of the resources and could not describe why they had not used them, other than general reasons such as time constraints. In some instances, practitioners who had not attended the initial training had not been made aware of the resources. Other practitioners said they had looked at the materials after receiving them, and sometimes said this had increased their knowledge in different areas. Others had used the resources more extensively, and found them useful. For instance, a couple of settings said they had 'dipped in and out', and engaged particularly with the maths resources, sometimes after discussions with the consultant. A practitioner in a setting in Region A said she sometimes glanced through the resources to get ideas about how to act on assessment data, but admitted she needed to use it more. Many had used the different grids to record assessment data, which will be discussed more in the section about collecting results.

Finally, AfC encouraged settings to get in touch by phone or email, or alternatively access the online forum whenever they needed support. However, these support mechanisms were not used, mostly because the settings did not feel like they needed it, or because they contacted their consultant directly. The online forum was not introduced to participants during the initial training day as the forum was not set up until later in the intervention. In addition, while the forum was meant to be available for all practitioners, it was clear that some lead practitioners had not made other practitioners within their school aware of the forum. All these factors could explain the lack of engagement with this particular resource, though interviews with practitioners would suggest they simply did not need any further support.

SMT support and involvement

Practitioners were also asked whether they had been supported by their senior leaders. The evaluation team also interviewed some senior leaders who were asked which support mechanisms they had provided. Most practitioners said the senior leaders were aware of the intervention, and said they would have been available to provide support, but that this had not been needed. As such, teaching staff had mostly been left to themselves to deliver the intervention. A typical comment was:

> *'She's been there to support, but we haven't really had many issues to ask her about. But she would find out an answer if we weren't sure, she's always been there to help.'* Nursery teacher, Setting 27, Region A

Some noted that implementation would have been easier if senior leaders had provided more staffing so the games could be played in quiet, separate rooms (see more detail in implementation section), but practitioners acknowledged this was not feasible. For some smaller providers, the senior leaders were sometimes practitioners themselves and were heavily involved in delivering the Toolbox. Some practitioners noted that senior leaders had supported them by giving them time and providing cover for the initial training day and consolidation days as well as training and support visits.

Summary: Did practitioners feel prepared to deliver the programme?

Overall, most practitioners said they felt prepared to deliver the programme and many emphasised that it was straightforward and self-explanatory. However, often this was said with reference to delivering the assessments themselves, that is, playing the apps. Generally, the support made practitioners confident in using the app but practitioners' preparedness to collect the assessment scores and subsequently inform and change planning and practice was less clear, and it varied between individual settings and practitioners.

More specifically, the initial training was seen as useful, particularly the opportunity to try it with children and to hear about the background and rationale for the intervention. Cascading the information from the initial training day to other practitioners were seen as fairly straightforward, and the main challenge was to make sure all practitioners used the games as an assessment tool rather than a teaching instrument. However, it should be noted that the cascading was fairly light-touch, mostly focused on playing the games, and it was apparent that non-attendees were sometimes less likely to know about other resources and less likely to understand that the assessment scores was supposed to inform planning and practice. The follow-up training worked well across all regions to address any start-up issues and providing reassurance that practitioners delivered the programme as intended. The support visits, however, varied fundamentally between settings, with some given more forward-looking support with ideas and advice about how to use the assessment scores to inform planning and practice.

Collection of assessment data

This section explains the process of collecting app assessment data and how settings experienced this part of the intervention.

Most practitioners found it easy to collect results, but there were quite different views among practitioners. The most common approach was to fill out the grid after the app assessment. Some used the sheets made by the delivery team while others seemed to use sheets made by their support officer. For the settings that found this easy, practitioners said they simply added the data into the spreadsheet after the child had played the app. For those who said it was time-consuming, there were strong concerns about the manual and the cumbersome nature of the data inputting. For instance, a teacher described the steps she had to go through every time a child had finished their game:

> *'What I've done is I've looked at the table at the end and then I've recorded the child's name and the age-related point on the table, say 1.8, and I've written it by their name … Then I've gone to photocopy them and then I've had to print it, I've had to load it onto my computer and then printed it from there, and then I've had to go and get the paper, bring it back and then try to match it to the numbers. And if you're doing that for twenty children …'* Teacher, Setting 13, Region B

It was typical among teachers to ask if it was possible to automate the process:

> *'I think it would be better if it went into, say, a database, and we just had access to it, and it was a really simple user interface.'* Teacher, Setting 18, Region B

A few settings did not seem to have collected the assessment data at all but instead relied on the practitioners' observations during the game, though this was not recorded.

Implementation and delivery

This section explores to what extent the intervention could be delivered within routine practice, including in relation to the expected frequency of assessments, number of children, time and space requirements, and other factors.

Frequency of assessments

The project specification required settings to use all three apps on a termly basis, that is, a baseline assessment in the autumn term (Term 1), one in the spring term (Term 2), and the final assessment in the summer term (Term 3).

The data collected through the apps was analysed with the aim of exploring actual use of the apps by settings. The app data suggests that settings, on average, used the apps less frequently than termly. The numeracy and vocabulary apps were more commonly used than the self-regulation app, and on average it appeared that settings had typically assessed children twice, rather than three times, over the course of the year, although there was variation by setting. However, it is important to note that there were notable inconsistencies in the ways in which identifiers for individual children had been entered into the apps and thus it is not possible to precisely identify the number of times the apps were used. It is possible, therefore, that these figures may understate the frequency of use (see Appendix 5 for further information on the use and analysis of the app data). It is also important to bear in mind that children may join and leave the setting throughout the course of the year, so some children may not have been present in the setting for the full year and therefore not able to be assessed at all time points.

The survey findings demonstrated mixed experiences among settings. A third of respondents agreed it was 'easy to fit Toolbox assessments into my usual teaching activities', a third of respondents were neutral and a final third disagreed with this statement. Of the 14 settings we interviewed towards the end of the intervention in June and July (Term 3), there were a number of settings that had not fully completed three rounds of termly assessments. In a few cases, this was merely due to a delayed start, while other settings had not been able to complete all the assessments as intended, or had deliberately chosen not to. These settings were:

- One setting in Region B decided it would not be useful to play any of the games in Term 3. The setting had not used the self-regulation app at all during the programme because it had judged it was too similar to its existing self-esteem indicator.
- One setting in Region C had been stretched in terms of staff ratios, especially due to a large intake of new children in January. As a result, it was only able to complete assessments during quieter moments in the nursery, and particularly during times when preschool staff had been able to support the nursery. This meant that the setting had not managed to complete assessments for all intended children in Term 2 and Term 3. Due to these staffing constraints, the nursery teacher said they had not 'really built [Toolbox] into our routine practice.'
- Some settings seemed to have used the apps more frequently due to misunderstandings about the intervention. For instance, one setting in Region B had used Toolbox as a termly assessment for individual children, but also seemed to have used it as an additional teaching tool, including for group activities in which the same children took turns to control the iPad. Similarly, a setting in Region A said it had let the children use it again when they had specifically asked if they could play. This interview was conducted during Term 2, so the issue may have been resolved during subsequent support visits. Similarly, one setting had used the apps too much in the beginning, but this was resolved quickly during the first support visit.
- One setting in Region B had experienced technical issues with its iPad, which caused a delayed start to the project. This meant it had only done two rounds of assessments, in Term 2 and Term 3.
- One setting in Region B did the first round of assessments with all 65 three- to four-years-olds at the end of Term 1 but decided not to complete any assessments in Term 2 as it was still analysing the results but then use the apps again in Term 3. This was attributed to the slightly delayed start in Term 1.
- The nursery manager in one setting had only done the self-regulation app in Term 1. She thought it was good to do the assessment as a baseline to see where children were on the scale but due to her strong knowledge of the children's abilities, she did not feel it would be useful to do the self-regulation app again.

Number of children

The participating settings varied considerably in size, thus it is not surprising that analysis of the app data shows notable variation in the number of games played. In some larger settings, it was agreed that the apps would just be used with one class or room.

Information on the approximate number of pupils was provided by AfC where this proved feasible, however, it was not possible to collect precise data on the number of children with whom the apps could potentially have been used. As noted earlier, the app data is also limited in that the identifiers do not necessarily identify children over time, so it is very difficult to explore with accuracy whether settings did assess all eligible children. Nevertheless, for a subset of around half the settings where some information on the number of children to be assessed was available, the app data suggests that typically most eligible children were assessed, even if not as frequently as expected.

For each app, a small number of children appeared to be younger than three or older than four, however the majority of children were aged either three or four. That said, information on age was also missing for a significant proportion of children (for example, for around 20% of children for the numeracy app). More broadly, resolving some of these issues relating to the entry of data into the app would be an important issue for consideration in future use of the app by practitioners as well as for any subsequent evaluation of efficacy (Appendix 5 provides further detail on how the app data was used and analysed as part of this pilot study).

Location

Settings were also asked about the location in which the apps were used. The programme specifications did not require settings to play the numeracy and vocabulary app in any particular location, but most settings preferred to play the apps in a separate, quiet room as the classroom was often described as being 'too distracting' due to the noise level, interruptions by other children, and disruptions when practitioners had to deal with incidents. Some practitioners were concerned that other children overheard answers on the vocabulary app, which could potentially affect their own subsequent assessment scores. However, staff ratios and sometimes room capacity meant that many settings were unable to complete the assessments in a separate, quiet room in the setting. Some tried to deal with the issue by using a quiet corner, a snack area, or designated technology areas in the classroom, or play the games inside or outside during free flow, though some practitioners noted that outside spaces were not ideal as the child wanted to run around with their friends. A practitioner who did not have access to any spare rooms had considered taking children into the baby room, but was concerned about the distraction from being in a different environment. Some practitioners thought it was fine to use the apps in the classroom and did not seem to consider the noise and distraction level as a significant issue.

Importantly, some settings had used mixed approaches throughout the programme and for individual children, depending on staff capacity on the specific day. This means that individual children are likely to have been exposed to different environments during different rounds of assessment. This could potentially impact on the accuracy of the assessment and progress scores. It was relatively common among practitioners to say that when apps were used in the classroom, they 'didn't feel it was giving a true picture of the children'. In fact, a nursery room leader explained that they had not been able to take the children out to a separate room during the assessments after Easter, and they had noticed lower than expected scores compared to previous assessment rounds, which they attributed to the distractions in the classroom environment.

Time

Settings were also asked whether there was a specific time of day or week that they usually used the numeracy and vocabulary apps with the children. The programme specifications did not require settings to use the apps at a particular time, and all settings responded that they used the apps whenever it fitted in to the daily routine, particularly based on whenever there was a practitioner or teacher free. Other specific times mentioned by practitioners included quiet times (such as the beginning or end of days), when other children were outside, or days where there were fewer children in the setting. Some practitioners also aimed for times where children were fresher or more enthusiastic, such as in the mornings or after they had eaten lunch. Some also said they used the apps whenever a child actively asked to do it, which was usually whenever another child had just done it. In some settings it varied because some children were only in the nursery in the morning or the afternoon. One setting in Region A had taken a more systematic approach during the last round of assessments where it had blocked out a week to carry out the assessments. Without this commitment, it felt the assessments would never have been completed as practitioners would always prioritise something else.

Length of assessments

There are two sources of information about how long it took to complete each app: (1) the app data includes information about how long each assessment took and (2) during the setting interviews practitioners were asked specifically about the length of time it took to complete each of the three apps. Most practitioners had not timed this in a systematic way, so most answers were based on estimates, and indeed our findings suggest practitioners may underestimate how long it took on average.

Table 6: App completion time, by data source (app data and estimates by practitioners)

App	App data: Mean	Practitioners: Mean	App data: Range	Practitioners: Range
Numeracy	10 mins	5–10 mins	0–17 mins	2–15 mins
Vocabulary	11 mins	3–5 mins	1–17 mins	2–10 mins
Self-regulation	13 mins	3–5 mins or 10–15 mins*	1–22 mins	N/A

*Depending on whether practitioners completed it alone or in pairs or groups.
Note: The app data completion times are based on data from all completed games in all settings. The mean estimates from practitioners are not collated and averaged in any systematic fashion. Instead, the mean is simply estimated based on the most common responses. The range is the lowest and highest estimate among all interviewed practitioners.

For the numeracy app, the app data suggested that the mean completion time was ten minutes, ranging between 0–17 minutes. The typical practitioner estimates seemed to be mostly between 5–10 minutes, but ranged from two minutes for low-achieving pupils to 15 minutes for high-achieving pupils. As discussed in the section on pupil engagement, the upper end was considered to be too long and future revised versions of the app should consider if there are ways to

reduce this time, such as using existing performance and age data to start the game at a more appropriate level for the individual child.

For the vocabulary app, the app data suggested that the mean completion time was 11 minutes, ranging between 1–17 minutes. Most practitioners said it took less time than the numeracy app. Their typical estimates seemed to be between 3–5 minutes, ranging between 2–10 minutes depending on the ability of the child.

For the self-regulation app, the app data suggested the mean completion time was 13 minutes, ranging from 1–22 minutes. Practitioners' self-reported responses varied substantially, probably reflecting the variety of the extent to which individual practitioners reflected on each question. For instance, one practitioner said:

> *'It depends how much I thought about it, I think; I think sometimes if you think too much you start doubting yourself whether you're right or not, so I tried to do it like really quick and not think too much.'* (Manager, Setting 22, Midlands)

Typically, participants said it took between 3–5 minutes per child, though others said it took between 10–15 minutes. The longer estimates were typically, but not always, made by those who did it in pairs or groups who noted that it was much more time-consuming than completing it individually.

It might also have mattered whether interviews were conducted in Wave 1 or Wave 2 of the evaluation as some respondents said they were completing the self-regulation questionnaire more quickly as they got used to the questions. Generally, some practitioners said it was difficult to interpret some of the questions due to how they were phrased, and one practitioner described them as 'unnecessarily complicated' which meant it was 'very easy to put down a wrong answer'. The only specific example given was that some questions were double-negative. In fact, one of the practitioners in Region C who had completed the app together with a colleague had initially done this to make sure she had understood and interpreted the questions correctly, and only then experienced that the discussions with the colleague was useful.

A few practitioners said they had to do the app in their own time as the length of the questionnaire simply made it impossible to fit into the normal day in the setting.

Were the apps only played during setting hours?

Settings were also asked about whether parents had been involved in the programme and whether they were aware of parents downloading and using the apps at home. Parental involvement was not part of the project design, and this question was asked to check that children had not been using the games at home, which would affect the validity of subsequent assessment results. Apart from the initial parent letter explaining the project, almost all practitioners said they were not aware of parental involvement, and they were not aware that any parents had purchased and downloaded the games at home. Only in two instances practitioners said that a parent had been interested in buying the app but they had no information about whether they had actually done it.

Instead, the findings from some settings highlighted that it might be beneficial for future rollouts to consider involving parents more intentionally, through feeding back the assessment results and making suggestions as to activities that parents can do with children at home in response to assessment scores. A manager in Region B said they shared the app results alongside other information in their termly meetings with parents. She said this had worked well as parents seemed to find it 'quite handy' to see the progress of the child, and in the one case where a child's progress had stalled, the parents had asked what they could do to help and the practitioner had given them some ideas for home activities. A nursery leader in Region B had a similar experience when she shared the results with parents. Some had been really interested and were provided with activities they could do at home to improve the child's ability in certain areas.

Children's enjoyment of using the apps

One of the intended intermediary outcomes in the theory of change model is that pupils enjoy using the apps. One of the key rationales for the Toolbox intervention is that the apps are an effective assessment tool, but children themselves don't realise that they are being assessed. Therefore, during case study interviews, settings were asked about the children's engagement and enjoyment with the vocabulary and numeracy app, respectively.

Broadly, practitioners responded that children really enjoyed playing the numeracy app and saw it as playing a game and as a fun activity rather than an assessment. This was generally highlighted as one of the most positive aspects of the numeracy app. The answers for the vocabulary app were more mixed, and typically practitioners pointed out that children preferred the numeracy app due to the interactive element and independency as children were in control of playing the numeracy game. The numeracy app was described as more interactive, more challenging and 'more than just saying a word'. This meant that sometimes children got a little bored with the vocabulary app, particularly the highest achieving children. One of the most negative practitioners commented:

> *'They go, "Oh, can we play?" and then when they come you say, "Well, no, don't touch the screen. It's just me that touches it," and they're like, "Oh right. I've had enough now."'* Reception teacher, Setting 1, Region A.

One practitioner in Region B, who did the apps consecutively with each child, said they always completed the vocabulary app first or else children would not maintain the concentration throughout both games.

But practitioners pointed out that the vocabulary assessment was relatively short, so the less interaction and engagement was not generally seen as a problem. Practitioners said that anything with an iPad usually made children engaged:

> *'Anything to do with an iPad, they like it.'* Setting 16, Region C, Nursery teacher.

> *'When they see the iPad coming out there's a big queue: "Can I go, can I go first, can I go first?"'* Preschool room leader, Setting 22, Region B.

The downside to the numeracy app was that assessments lasted for a long time for high-achieving children. Practitioners invariably said that 'it just kept going' and some estimated it could last up to 15 minutes. Practitioners pointed out that this was a long time for a nursery child, and the bright children often got bored and started to get distracted before they got to their 'working age', leading teachers to question the accuracy of the subsequent assessment data as the app would underestimate the level and progress of these children. For instance, a nursery teacher in Region A had observed that high-performing children had not achieved much progress, and attributed it to this problem. A nursery teacher in Region B pointed out that their existing assessment tool, WellComm, starts slightly below children's expected ability, based on their age and previous performances. She recommended that the Toolbox app was revised in a similar fashion so children start at a more suitable level. Alternatively, a nursery teacher in Region A suggested they should have the option to pause and return at another time.

There were also some practitioners who pointed out that the apps lacked variety. For instance, one said some children quickly got bored with similar counting exercises in the numeracy app and stopped counting carefully as a result. Others said the vocabulary app used the same prompt every time (for example, 'How else would you say that?') and it used the same photos and words every time. With respect to the numeracy app, many also noted how children were not patient enough to wait until the speaker had finished the question, and some noted that the questions were slightly long-winded at times leading some children to 'lose their train of thought' by the time they were allowed to answer.

Some practitioners said that five incorrect consecutive answers seemed to be too high a threshold for ending the numeracy game. As children reached beyond their actual level and started guessing, they were sometimes lucky in getting one answer correct once in a while. This meant pupils sometimes continued for a while beyond their actual ability, which led to disengagement. Some suggested adding another exclusion criterion such as getting a certain number out of the last ten correct.

Facilitators and barriers to the programme

The delivery and evaluation of the pilot programme has highlighted a number of facilitators and barriers to successful implementation. Broadly, they can be categorised into six different areas, which are covered in the subsequent sections.

Time for implementation

When practitioners were asked about potential facilitators or barriers, many mentioned 'time' as the main barrier. There are four components of the intervention that required time for practitioners:

- time for training sessions, the consolidation event, and support visits;
- time to complete apps;
- time to collect assessment data; and
- time to act on learning from assessments.

Generally, practitioners seemed to believe the time required for attending training sessions, the consolidation event, and support visits was appropriate and feasible. However, some settings seemed to struggle with the time required to complete the apps, particularly for larger settings with many learners. The self-regulation app, in particular, was highlighted as time-consuming, particularly for those who did it in pairs or groups, and some teachers had to complete the questionnaire at home in their own time. A specific issue that was highlighted by some practitioners was the purpose of the 'specify button'. They said it took quite a lot of time to enter and sometimes disrupted the rhythm of the game, and children got bored while waiting for practitioners to type it in. These practitioners wondered what the data would be used for and only wanted to invest the time required to do this task if it was actually used for something useful that would inform future versions of the app, or be used to adjust the assessment scores.

In terms of collecting assessment data, it seemed varied to what extent settings considered time as an issue, but it was clear that at least for some settings it was a manual and very cumbersome system, which could be automated and shortened. Finally, some settings and practitioners felt they had not invested enough effort into acting on the learnings from assessments. To the extent this was due to a lack of time, earlier support during the initial training day as well as better targeted materials and support would reduce the time required to inform planning and practice.

Level of support

The evaluation identified that the level of support was an important enabling factor that determined to what extent practitioners collected the assessment data and acted on learning from assessment. For instance, as discussed previously, the support visits in Region B had provided more forward-looking advice and conversations about how to inform planning and practice, which in many cases had resulted in more instances of these practices.

Factors affecting practitioner engagement

The evaluation also identified a number of factors that seemed to affect practitioner engagement. Firstly, the different amount of information given to those practitioners who attended the initial training day and those who did not—particularly on the background, purpose, and rationale of the intervention—seemed to lead to different levels of engagement in the intervention, different levels of knowledge about the purposes of the intervention—in particular, the aim to inform practices and planning rather than collecting just another set of data—and different levels of understanding behind the rationale of the apps, in particular that the app is proven to be as effective as more complex and time-consuming assessment tools.

Secondly, the senior leaders who had typically signed up settings for the Toolbox intervention had rarely communicated properly with staff about the sign-up, including about the reasons for participating in the intervention and about existing assessment practices and priorities. This meant that Toolbox, in some instances, did not have the required buy-in from practitioners. For instance, some practitioners said one of the key focus areas of their setting was to 'de-technologise' their practice. They pointed out that children already play a lot of iPad games at home and said the setting's current priority was to reduce screen time in order to improve areas such as fine-motoring and writing with a pen. Other practitioners, with specific reference to the numeracy app, said they preferred doing those type of exercises with real objects in everyday practice, in interaction with an adult. In these settings, the practitioners had typically not been involved in the initial decision to sign up for the intervention, so this issue could be resolved by making sure that senior leaders in settings properly consult their staff before signing up for any future rollout of Toolbox. Similarly, in some settings, key members of staff were happy with the existing assessment practices, including specific assessment tools or focus on observational data. Finally, in some settings, key members of staff did not believe in the value of assessment overall, and had not been convinced by the presentation on the initial training day.

Factors affecting how apps are played in a setting

The physical environment was also identified as a facilitator or barrier in settings. As discussed previously in the report, staff ratios and space requirement meant that in some settings practitioners were able to play the games with pupils in a quiet environment without distractions, while other settings had to play the games in the classroom with distractions. In addition, some settings had limited access to WiFi, which made the data collection process more time-consuming as data was harder to upload.

Factors affecting perceived accuracy of data

A significant facilitator, or barrier, was the extent to which practitioners believed in the accuracy of the assessment data. If practitioners observed issues during the app assessments, which made them distrust the assessment data, then this made practitioners less likely to use the data to inform planning and practice. While practitioners generally trusted the data, and often said it confirmed their casual observations of the pupils' ability and progress, the evaluation identified that two aspects made practitioners less confident about the accuracy of the data.

Firstly, as previously discussed, practitioners were concerned about the long duration of the numeracy assessments for higher-ability pupils who sometimes lost engagement and their concentration by the time they reached their 'working age'. This meant that some practitioners thought the app underestimated the level and progress of higher-ability pupils.

Secondly, a number of gameplay issues led to concerns about the accuracy of children's assessment scores. For the numeracy app, two issues were mentioned by almost all practitioners. First, children often tried to 'click-and-drag' objects before the speaker had finished the question. Because this was not possible, it led children to believe that their original answer was wrong and would then go for another answer option by the time they were allowed to answer. Second, children were asked to count fruits such as bananas and tomatoes in very small, unstructured and overlapping bundles. At this age, the inability to do this was not seen as reflective of their counting abilities. The unstructured bundles also made it difficult to identify pupils' ability to count quickly using different shortcuts as there was no identifiable pattern. A nursery manager commented:

'At that age, really and truly you need to give them a good opportunity to be able to differentiate the amount of objects because that's part of what they need to do in their development.' Nursery manager, Setting 10, Region C

Other responses from practitioners focused on small design aspects of the app where children answered questions incorrectly because they were confused. It was felt that children could cope with the numeracy tasks but were hampered by design flaws. The following issues were raised:

- While pupils might have understood the numeracy concept, they were sometimes confused about the meaning of the questions mark in the place of a missing number.
- When asked about tallest and shortest, for instance, some pupils looked at the people or the buildings rather than the number.
- There was some confusion with full and empty jugs as some children simply thought some jugs were filled up with water and milk, respectively.
- In questions with dogs, some children found it confusing that they were all different breeds.
- Some practitioners thought the screen was too busy, at times leading children to focus on the wrong things.
- Some said the app lacked continuity without a logical flow between questions; others said questions and the accompanying screen were sometimes too similar to the previous one, leading children to assume it was the same question while in reality they had to point to the tallest rather than the shortest, for instance.
- One practitioner said that some children sometimes accidentally pressed the 'turn page' button rather than clicking on the robot that repeated the question.

For the vocabulary app, practitioners also identified a number of gameplay issues that led to concerns about the accuracy of the assessment data. The most common concern was that some of the photos were 'obscure' or 'deceiving'. Many suggested using real photos rather than 'silly cartoon photos', and some added that their setting was very big on using real photos in their teaching. A setting manager gave the example of the purple spider: 'They were all, like, "spiders aren't purple", they don't understand why it's purple' (Manager and practitioner, Setting 19, Region B). Some practitioners also said that photos were sometimes not specific enough in the sense that they were part of a scene which made it unclear for the children what word they were supposed to say. For instance, some said that children sometimes answered 'grass' instead of 'bridge', 'water' instead of 'bucket', and 'night-time' instead of 'snow'. Some specific pictures were mentioned repeatedly by practitioners, such as the feather that looked like a leaf and the cactus and the fountain, which for most children were not everyday objects and not something they had come across before.

Another common response was that the actions, such as winking, coughing, sneezing, sweating, and yawning, were difficult to understand for the children, simply because the practitioners perceived that the animations were poorly done. On the other hand, a couple of practitioners said the children's confusion about the actions were eye-opening and they had noticed that children were not actually using action verbs in their speech. A nursery teacher gave the example that when someone was yawning, they would say 'he is tired' rather than 'he is yawning'. A few practitioners said the children found those hilarious. A couple of respondents also noted that it took too long from the moment the speaker asks 'What is this?' to when the action actually happens. This meant that the child had sometimes already answered, and might not have paid attention when the action happened. Some also said that specific words seemed to favour specific child backgrounds, for instance the focus on musical instruments such as guitar and violin. Some practitioners noted that some children tended to talk around the photos, often using very beautiful vocabulary, but because they did not say the exact word, they often got a bad score, which in some instances was not considered to reflect their vocabulary and language ability.

Some practitioners also said children sometimes got confused with the speaker, both in terms of the speaker's phrasing of questions and the Australian accent which was described by some practitioners as 'off-putting' and unsuitable for U.K. settings. A few practitioners noted that some children often asked the practitioner to repeat or rephrase the question.

Thirdly, and finally, the location in which the app was played could affect the child's performance in the numeracy and the vocabulary game. As previously discussed, some practitioners had noticed that pupils scored better in quiet locations but it was not always possible to find suitable, separate rooms in settings. As such, some pupils' scores were not obtained under ideal conditions, and sometimes the location varied across terms for each individual child, which would affect the accuracy of the progress scores.

Factors affecting how assessment data is used

The evaluation identified a number of factors that affected how practitioners made use of the assessment data and especially if they used it to inform planning and practice. The support from the consultant during support visits was instrumental in making practitioners think about next steps, such as how they could inform their planning and practice and target specific children or focus on particular aspects of learning in the setting. This was particularly clear in Region B in which the consultant successfully made practitioners and settings reflect on the app assessment data, and how it could be used going forward. It was clear that practitioners' understanding of the key elements and objectives of the

programme affected how they used the assessment data. As such, some practitioners—particularly those who had not attended the initial training day and those who did not have a consultant who emphasised this aspect of the intervention—did not seem to use the data to inform planning and practice.

It was also clear from our evaluation that the engagement with the Toolbox app, including how and to what extent the assessment data was used, depended upon existing assessment practices in the setting. Sometimes, but not always, this varied between types of settings. For instance, independent and private voluntary settings often had no assessment practices in place and as such Toolbox offered a valuable and simple assessment tool. In contrast, school settings typically already had more complex assessment practices in place, which they were reluctant to replace.

Readiness for trial

This section assesses whether the programme is ready to be trialled at a larger scale, including consideration of the extent to which the training and resources make up a standard intervention which is replicable and scalable to a wider context beyond the pilot stage.

Specifically, the section is structured as follows:

- first, we outline formative findings, focusing particularly on those related to readiness for trial;
- second, we explore to what extent the training and resources constitute a standard intervention that can be replicated and scaled up, including whether the training and support contained the same identifiable components across settings and regions; and
- third, we describe the costs of the intervention.

Formative findings

Using Toolbox with other assessments

The intervention generally appeared to be of most benefit to staff at settings that did not already have comprehensive assessment procedures in place, with school settings more likely to have such processes. It would be useful to consider more carefully how Toolbox should be used, if at all, in conjunction with existing assessment tools. One of Toolbox's selling points is that it provides a simple assessment tool that is as effective as more complex ones. However, the delivery team should be clearer about the purpose of Toolbox when it is used together with other assessment tools. This could be taken into account during further roll-out of the intervention in future, for example through discussion with settings during the recruitment stage and initial training. The finding is important for any subsequent efficacy trial, which should be delivered to the ideal population and in the ideal settings. The finding suggests that the 'ideal setting' may depend on assessment procedures that are (or are not) in place or setting type, which could constitute inclusion/exclusion criteria for recruitment at efficacy stage.

Using and improving the apps

Some findings relate specifically to the apps themselves. These included the gameplay issues mentioned earlier, some of which were addressed during the course of the pilot. Some practitioners felt that the vocabulary app would have been more useful if it provided more information regarding learning sub-areas in a similar way to the numeracy app.

Practitioners needed more support and resources in how they used the assessment scores to inform practice, especially on what to do next after discovering gaps or low performance. The delivery team should consider introducing specific supporting materials at an early stage, and provide guidance on how they can be used, so practitioners consider how to use assessment data to inform planning and practice as early as possible. Resources were developed during the pilot, but some practitioners were unaware of these so these would need to be embedded and cascaded from the start of any project for future roll-out.

Practitioners responded to the findings from the Toolbox assessments in a number of different ways; it would be helpful, therefore, to provide guidance and resources on applying this information to group practices and to dealing with individual children. The reported changes in how practitioners worked with children tended to be in terms of individual activities rather than changes to their overall teaching approach so it could be helpful to focus on this. It is also worth bearing in mind that the settings that made changes tended to be non-school settings. The numeracy app led to the most reported changes in activities and overall approaches, so this area might be particularly beneficial to have guidance and activities for.

Considerations for future trials

The main aspects of the intervention to be developed before any future trial are as follows:

- The approach to guidance offered during support visits, including provision of supporting materials, should be as consistent as possible across different regions. This will be particularly important in a potential larger rollout

of the intervention where more consultants are likely to be associated with the project. Any future evaluation should observe these training sessions.

- The intervention should provide more guidance and materials to help lead practitioners cascade training to colleagues following the initial training day. For example, this could include a greater focus on this element in the initial training or a manual to use back at the settings.
- Making the intervention more time-efficient would increase its value to practitioners, including by automating data collection, shortening the numeracy app for high-achieving pupils, and reducing preparation time for practitioners by providing more support and targeted materials on how to use assessment data to inform planning and practice.
- The recruitment of settings for any future trial requires careful thought to ensure that senior leaders and practitioners have considered how Toolbox would complement existing assessment practices, and whether the setting infrastructure, staff numbers, and other teaching priorities are likely to provide a conducive environment for the intervention.

Replicability and scalability

This section assesses whether the programme is ready to be trialled at a larger scale, that is, we consider the extent to which the training and resources make up a standard intervention that is replicable and scalable to a wider context beyond the pilot stage.

In a larger trial, the delivery team would plan to implement Toolbox on a similar basis to the pilot. The evaluation findings suggest that this is possible. The delivery team has indicated that Professor Melhuish is able to deliver initial training sessions as in the pilot intervention, including running a higher number of training days as a larger trial would be expected to include more regional clusters. Similarly, a larger trial would potentially need more than one consolidation event at the end of the intervention. This participation of one of the co-founders of the Toolbox app was identified as important in the trial evaluation as it increased buy-in from practitioners to hear about the rationale and objectives of Toolbox. However, the involvement of one specific individual may not be sustainable on a permanent basis for potential subsequent RCTs and future rollouts. While the involvement of Professor Melhuish was seen as important, it is, however, not essential for the initial training days to be effective. The evaluation team's judgement is that other trainers, for instance consultants who have done support visits in previous rollouts, could run the initial training day if this project was on a larger scale.

The delivery team would also need to give further consideration to the optimal number of consultants. As there would be a larger number of settings, the delivery team would need to consider the trade-offs between increasing the workload of each individual—that is, increasing the number of settings they cover including potential longer travel to settings—compared to increasing the number of regional clusters with potential impact on the consistency across areas and settings.

The pilot evaluation identified an issue with consistency in the support provided by consultants across regions. This would need to be addressed to make sure the support contains the same identifiable components across regions so child app data can be meaningfully compared. To ensure consistency across areas and individual settings, the evaluation findings suggest that Professor Melhuish and the delivery team should train future consultants; such training should cover the purposes and objectives of the intervention as well as how to conduct support visits and use materials in the settings.

The online forum was scarcely used by practitioners during the pilot. However, they were not introduced to this resource during the initial training day as it was not launched at the time (this had not originally been planned to be part of the intervention). It may prove more useful in a larger trial where it would be introduced from the start. The forum could also be a platform for sharing experiences of the programme, including asking settings to upload their own resources and activities that had been created as part of Toolbox.

Other necessary and potential changes, which were suggested in the formative findings section, should not affect the scalability of the intervention. Finally, some changes were made during the course of the pilot and therefore would already be in place for the start of a larger trial, such as guidance on additional resources, data collection sheets, and improvements to the app layout.

Cost of intervention

In this section we discuss the cost of the intervention to settings. We include direct costs (such as the cost of training), indirect costs (for example, any interventions or materials bought as a result of the assessments), as well as costs in terms of staff time required.

The cost estimates are largely based on information provided by the delivery team. Information on costs was discussed during the final IDEA workshop, focusing particularly on identifying the different components of the costs involved. As

far as possible, costs are estimated based on the costs as delivered in the pilot and are based on the cost from the perspective of participating settings. Issues relating to costs and time incurred were also explored as part of the interviews with settings and the follow-up survey included questions on additional costs and staff time required (these two questions were asked of managers and deputy managers only). Very few respondents to the survey indicated additional costs and where these did so they generally related to time costs, plus one respondent noted photocopying costs were incurred in order to store results.

In line with EEF guidance, we describe costs below in terms of direct marginal costs, prerequisites, and staff time.

Direct, marginal costs

These costs are directly attributable to settings' participation in the programme. In the case of the Early Years Toolbox pilot, these costs relate to:

- the provision of training (initial and follow-up);
- travel for practitioners to attend initial training;
- ongoing support (visits) and the consolidation event;
- cost of apps
- cost of iPads—although note discussion below regarding whether this could be considered a prerequisite; and
- the provision of supporting materials (additional resources).

The main cost of providing the initial training relates to the time required by Professor Melhuish, the delivery team, and the consultant for the region. This would include both their time in delivering the training sessions as well as preparation time and any other costs incurred such as travel of the team. All venues were provided free-of-charge by the host settings during the pilot. The total cost of delivering all three initial training sessions was estimated at £9,690 (equivalent to £3,230 per training day).

Settings were given an allowance of £550 per setting to cover both travel costs and cover for attending the initial training. As cover is considered in terms of time rather than a financial cost, we assume that 30% of this amount is used to fund travel costs in our estimates below.[9]

The main component of cost for the follow-up training and support visits is also the time of the delivery team and consultants. This was estimated at an average of £700 per visit. The total cost of the consolidation event towards the end of the programme was £2,877 (covering room hire, equipment, and refreshments).

Direct costs also include the purchase of the apps themselves (the self-regulation app is available free of charge but there is a small charge for the numeracy and vocabulary apps). The cost estimates also include the cost of an iPad in order to use the apps. In the pilot this also included provision of technical support with the iPad by Apple, which may not be required in a future rollout.

Other direct costs involved, but for which we do not have estimates, relate to:

- providing additional materials to settings—the main cost here is likely to be the time of the delivery team in compiling the resources; and
- providing the online forum.

Finally, although this would not count as a direct cost, it is worth noting that indirect costs may be incurred by settings if they decide to purchase additional resources or materials following on from using the apps. It would be debatable whether such costs could truly be attributable to the programme (for example, settings may well make such decisions based on a range of factors) and would be likely to be very variable, and so these are not factored into estimates of the cost of the intervention.

Prerequisites

A prerequisite for using the programme is access to the internet; it is assumed that this would be available in settings and thus is treated as a prerequisite. The programme also requires the setting to have an iPad. This has been included as a direct cost for the purpose of calculating the cost estimates, and was provided to settings in the pilot, however, this could potentially be considered as a prerequisite for a future trial.

Setting staff time

Staff time includes time required for training, time spent delivering the intervention, and time required to act on the results.

[9] On the basis of an approximate estimate by the delivery team.

The initial training sessions required around one day of time from participants (once travel time is factored in), with most settings sending two practitioners to attend. Time is also required for practitioners who attended the initial training to cascade the training to colleagues, as well as time from staff to attend the follow-up training and the support visits from the consultants (three support visits in total).

The intervention is designed to be used within the normal setting day. As discussed earlier in this report, in a few cases, practitioners reported that they had completed the self-regulation app in their own time but this is not the expectation of the programme. On average, based on the app data, the vocabulary app took 11 minutes to complete per child, the numeracy app took 10 minutes, and the self-regulation app took 13 minutes (although practitioners generally estimated a shorter average length of completion). The intention was for each app to be used with each child once per term (three times over the course of the intervention or per academic year). Time is also required in order to record the results from the assessments.

The intervention also requires time to act on learning from the assessments, however this is very difficult to quantify and may well be very varied.

In terms of potential staff cover time required, this would comprise time to attend the initial training as well as to participate in the follow-up and support visits.

The follow-up survey asked respondents approximately how much staff time had been spent on the intervention.[10] Only a few responses were received, and these were varied, and with some respondents only mentioning, for example, the time spent on training and support visits.

We use the information provided by the delivery team on direct marginal costs to construct an approximate estimate of the cost per child per year (Table 7). Although we do not have precise information on the number of pupils in participating settings, we assume an average of 27 pupils per setting.[11] All costs can effectively be considered as start-up costs, so the average cost per pupil over three years is equivalent to around £60. This means that the cost of the intervention would be considered 'very low' based on standard EEF cost ratings.

Table 7: Cost of delivering Early Years Toolbox

Item	Type of cost	Cost	Total cost over 3 years	Total cost per pupil per year over 3 years
Provision of initial training	Start-up cost (total for all settings)	£9,690	£3,230	£13
Follow-up training/support visits	Start-up cost per setting (follow-up training plus 3 support visits)	£2,800	£933	£35
Travel costs (to attend initial training)	Start-up cost per setting	£165	£55	£2
IPad (including support) and apps	Start-up cost per setting	£722	£241	£9
Consolidation event	Start-up cost (total for all settings)	£2,877	£959	£1
Total				£60

Note: costs are rounded to the nearest £. For the estimates of total cost over 3 years (fourth column), the costs for the provision of initial training, and for the consolidation event, are the cost for all participating settings. All other costs in this column present the cost per setting.

While the costs presented above are based on the design of the programme as in the pilot study, as the intention is that any future trial would adopt a similar approach to implementation, the cost components should remain relatively similar. Implementing on a larger scale is, however, likely to necessitate some changes, such as holding multiple consolidation events rather than one for all participating settings, which would have some implications also for costs.

[10] The relevant question was: Approximately how much staff time has been spent on the Early Years Toolbox and associated activities since your setting started participating in the programme (including attending training and/or taking part in any follow-up support visits)?

[11] This was the average size of settings assumed in the study plan.

Conclusion

Summary of pilot findings

Question	Finding	Comment
Is there evidence to support the theory of change?	***Mixed***	***There is some evidence that the intervention improved practice and changed practitioners' behaviour, but only limited evidence on whether the intervention enhanced practitioner understanding of child development, ability to target support, and whether it could improve child outcomes. However, this evidence is based upon staff report that may under or overestimate change.***
Was the approach feasible?	***Yes***	***The approach is feasible and could be improved by automating the presentation of assessment data and making other improvements to gameplay to help increase take-up.***
Is the approach ready to be evaluated in a trial?	***Mixed***	***EYT is now more replicable and scalable but consistency in the cascading of training and consultant delivery is needed to ensure a future trial has the best chance of showing impact.*** ***More support is required so that practitioners know how to respond after gaps in children's competencies are identified.***

Formative findings

Several refinements were made to the intervention during the course of the pilot. These would therefore already be in place at the start of the rollout of any future trial. In addition, as discussed earlier in this report, the evaluation identified some other ways in which the intervention could be further improved while recognising that there may be practical or financial constraints in doing so. We summarise these below, including highlighting points of relevance for any future trial.

Using Toolbox with other assessments

- The intervention generally appeared to be of most benefit to staff at settings that did not already have comprehensive assessment procedures already in place (school settings were more likely to have such processes). The delivery team should consider more carefully how Toolbox should be used, if at all, in conjunction with existing assessment tools. One of Toolbox's selling points is that it can provide a simple assessment tool that is as effective as more complex ones. However, the delivery team should be clearer about the purpose of Toolbox when it is used together with other assessment tools. This could be taken into account during further roll-out of the intervention in future, for example through discussion with settings during the recruitment stage and initial training.
- It is also noteworthy that some practitioners saw the Toolbox as an 'objective' tool for verifying other assessment methods such as observations. This could be taken into account when supporting or promoting the Toolbox.

Using and improving the vocabulary app

- Some practitioners felt that the vocabulary app would have been more useful had it provided more information regarding learning sub-areas in a similar way to the numeracy app, and a small number of practitioners compared it to other vocabulary tools that they felt provided more detail, such as WellComm. It may be useful to consider whether the vocabulary app might incorporate such information in future and/or provide more guidance on how findings from the vocabulary app could be used to inform teaching.
- A couple of practitioners had used the vocabulary app to identify speech and language support needs, using the findings from the app as evidence. It may be helpful to provide some guidance or support around clear and appropriate ways to use the app data in this context.

Using and improving the numeracy app

- Improvements to the numeracy app include resolving some of the identified gameplay issues, potentially shortening the numeracy app for high-achieving pupils (potentially by adjusting the start point to an appropriate level), and providing additional exclusion criteria.

Using and improving the self-regulation app

- Staff at some settings completed the self-regulation app as pairs or groups so it might be useful to provide guidance around this. Also, some staff observed that areas covered by the app overlapped with the Early Years Foundation Stage (EYFS) profile so some guidance about how findings from the app can be applied to the EYFS could be useful.

Using the results from the apps in practice

- Practitioners felt that follow-up guidance on what to do with assessment results for all children would be useful. Resources were developed during the pilot, but some practitioners were unaware of these so these would need to be embedded and cascaded from the start of any project for future roll-out. The approach to guidance in support visits should also be as consistent as possible across different regions.
- Several staff felt there was a need for more support and resources on what to do next after discovering gaps or low performance. A couple of staff compared the Toolbox to the vocabulary tool WellComm, which they reported provides ideas for next steps. The delivery team should consider introducing specific supporting materials at an early stage, and provide guidance on how they can be used, so practitioners consider how to use assessment data to inform planning and practice as early as possible.
- Some practitioners found the Toolbox apps useful for identifying areas where they themselves wanted to learn more, so perhaps further strategies and resources could be developed to support this.
- Practitioners reported a mix of approaches to responding to findings from the Toolbox assessments so it would be helpful to consider providing advice or resources on a range of tasks for both groups and also individual children. The reported changes in how practitioners worked with children tended to be in terms of individual activities rather than changes to their overall teaching approach so it could be helpful to focus on this. It is also worth bearing in mind that the settings that made changes tended to be non-school settings.
- The numeracy app led to the most reported change in activities and overall approaches, so this area might be particularly beneficial to have guidance and activities for.
- Staff at a couple of settings had decided to use the Toolbox apps as a teaching tool rather than for assessment and some practitioners questioned whether the time spent conducting assessments with the Toolbox could be better spent teaching or working with children one-on-one. This suggests that it may be helpful to review communication, specifically, by outlining how using the assessments can be of benefit to both children and teachers.

Training and support

- The intervention should provide more guidance and materials to help lead practitioners cascade training to other practitioners following the initial training day. For example, this could include a greater focus on this element in the initial training or a manual to use back at the settings.
- The delivery team should make sure that the training of consultants is improved so the support visits are more effective and consistent across all regions. This will be particularly important in a potential larger rollout of the intervention where more consultants are likely to be associated with the project. Any future evaluation should observe these training sessions.

Considerations for future trials

- Making the intervention more time-efficient would increase its value to practitioners, including by automating data collection, shortening the numeracy app for high-achieving pupils, and reducing preparation time for practitioners by providing more support and targeted materials on how to use assessment data to inform planning and practice.
- The recruitment of settings for any future trial requires careful thought to ensure that senior leaders and practitioners have considered how Toolbox would complement existing assessment practices and whether the setting infrastructure, staff numbers, and other teaching priorities are likely to provide a conducive environment for the intervention.
- Further development is needed to consider how the apps can be developed to consistently identify children over time. At present, this is reliant on the practitioner entering a consistent reference for children each time they use the app. If processes for collecting and reporting the app data are to become more automated, the use of consistent identifiers would be essential in order to do this with accuracy.

Interpretation

Evidence of promise

The theory of change model posits that shorter-term outcomes from the intervention will focus on improving knowledge and understanding, practice and behaviour among practitioners, and then, in the longer-term, lead to improvements in children's outcomes.

The findings from the pilot find some evidence to support the potential for the intervention to change practitioners' practice. Results from the survey indicated the majority of respondents felt the Toolbox had helped improve their practice. Findings from the interviews generally identified limited change in practitioners' understanding of children's skills and competencies, but with some reporting more substantial impacts. The most impact in this respect typically related to the numeracy app.

The pilot did not aim to establish impacts on children's attainment, other than through asking practitioners whether they felt the intervention was likely to bring about benefits in this respect. These views were mixed, with some reporting positive benefits, typically for numeracy or vocabulary, but others indicating no effects as yet. However, a greater emphasis on what practitioners may do in response to the results of the apps, and greater support to achieve this, would be likely to increase the chances of the intervention having impact.

Feasibility

In general, the intervention was feasible for practitioners to implement, although some settings did not manage to complete all the assessments or needed to undertake some outside of the normal setting day (in the case of the self-regulation app). Time was the most common barrier to using the app, therefore means to make the intervention more time-efficient wherever possible would be valuable. In particular, automation of data collection and reporting to provide assessments results in a user-friendly manner for practitioners would be welcome. This would be likely to require further development work on the app in order to create a user-friendly interface for this purpose.

Readiness for trial

Refinements were made to the intervention throughout the course of the pilot and the programme should be largely replicable and scalable. Some changes would be needed to operate on a larger scale, such as increasing the number of regional training days, however, the developer has indicated it would be feasible for it to accommodate this increase. As previously discussed, this would also involve decisions about the optimal number of consultants required, with thought given as to the balance between achieving consistency across consultants in their approach whilst maintaining manageable workloads for consultants. More structured guidance on cascading training to colleagues would be valuable to help to better embed the programme. The monetary cost of the intervention to settings is likely to be fairly low, thus time involved is more likely to prove a barrier than financial cost.

Limitations

As with all research studies, the evaluation has some limitations. Practitioners from all pilot settings were interviewed as part of the evaluation. While these include a range of settings of different types and characteristics, the small number of settings overall means that generalisability of results beyond the pilot is limited. We cannot rule out the possibility that those settings taking part may well be those that are particularly engaged and motivated to bring about change in their settings. Furthermore, as stated throughout the report, the survey findings are based on a relatively small number of respondents, and there may well be some bias in the types of individuals who choose to respond.

The pilot did not set out to quantitatively assess the impact of the intervention on child outcomes. While data from the apps was collected, this was primarily with the aim of establishing usage rather than outcomes as there is no control group within the pilot in order to draw comparisons. In practice, analysis of the app data was also limited by difficulties in consistently identifying children over time. Nevertheless, this has highlighted a useful point for future development.

Future research and publications

Any future randomised controlled trial (RCT) to evaluate the impact of the Early Years Toolbox would need to take into account a number of factors.

The nature of the intervention means that any future trial would be likely to require setting-level randomisation. The learning and knowledge that practitioners gain during the course of the intervention may well result in changes being made not just for individual pupils, but also at a whole setting level. Randomising nursery classes or rooms within settings may also be difficult in a practical sense if practitioners work across these groups. Any randomisation below setting level would be likely to entail significant risk of contamination.

As in any trial, the choice of outcome measure(s) would also require careful consideration. As the Toolbox itself is inherent to the programme, an alternative outcome measure is likely to be required in order to consider impacts, at least as the primary outcome. Given that the most positive feedback in the pilot typically related to the numeracy app, it may be worth focusing on a numeracy measure as the primary outcome. This was also an area often felt to be one in which practitioners particularly welcomed further support in improving their practice. Potential impacts on literacy and self-regulation could also be considered as secondary outcomes. In choosing an appropriate outcome measure, it will be necessary to consider not only reliability and robustness, but also to be mindful of potential burdens on settings. In the longer-term, it may also be valuable to use administrative data on outcomes for children through national assessments such as the new reception baseline assessment and/or the Early Years Foundation Stage Profile.

In addition to outcomes for children, it may also be useful for a future RCT to consider the impacts on practitioners, such as their confidence and views on assessing children.

Other factors to be considered include the timeframe over which the intervention should best be tested. In the pilot, the intervention ran for one academic year. However, it is possible that any impacts on children's outcomes may take longer to emerge as it may take time for changes in practice to become embedded and to then have an impact on children's outcomes.

The key research question to be addressed by such a trial would be whether the intervention has an impact on children's outcomes. Given the differences in views among different types of settings demonstrated in the pilot study, it may also be of value to explore whether the intervention has a differential impact in PVI settings compared with school-based settings. It may also be worth exploring, therefore, whether the trial should be focused on a particular type of setting, although the consequences for recruitment and implementation of the trial should be carefully considered. One further practical consideration will be to ensure that as part of any recruitment process, as far as possible, that participating in the trial is not a decision taken solely by senior leaders but is also discussed with the practitioners that would be ultimately responsible for delivery.

References

Connor, C. M., Morrison, F. J., Fishman, B. J., Schatschneider, C. and Underwood, P. (2007) 'The Early Years: Algorithm-Guided Individualized Reading Instruction', *Science*, 315, pp. 464–465. DOI: 10.1126/science.1134513

Connor, C. M., Morrison, F. J., Schatschneider, C., Toste, J., Lundblom, E. G., Crowe, E. and Fishman, B. (2011) 'Effective Classroom Instruction: Implications of Child Characteristic by Instruction Interactions on First Graders' Word Reading Achievement', *Journal of Research on Educational Effectiveness*, 4, pp. 173–207.

Howard, S. and Melhuish, E. (2016) 'An Early Years Toolbox for Assessing Early Executive Function, Language, Self-Regulation, and Social Development: Validity, Reliability, and Preliminary Norms', *Journal of Psychoeducational Assessment*. DOI: 10.1177/0734282916633009

Howard, S, Melhuish, E. and Chadwick, S. (2019) Early Years Toolbox website: 'About the Toolbox'. **http://www.eytoolbox.com.au/about**

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., et al. (2014) 'Better Reporting of Interventions: Template for Intervention Description and Replication (Tidier) Checklist and Guide', *BMJ*, 348, g1687. DOI: https://doi.org/10.1136/bmj.g1687

Humphrey, N., Lendrum, A., Ashworth, E., Frearson, K., Buck, R. and Kerr, K. (2016) 'Implementation and Process Evaluation (IPE) for Interventions in Education Settings: An Introductory Handbook', London: Education Endowment Foundation. **https://educationendowmentfoundation.org.uk/public/files/Evaluation/Setting_up_an_Evaluation/IPE_Guidance_Final.pdf**

NatCen (no date) 'Study of Early Education and Development (SEED)'. **http://www.seed.natcen.ac.uk/**

York, B. N. (2014) 'Know the Child: The Importance of Teacher Knowledge of Individual Students' Skills (KISS)', Center for Education Policy Analysis (CEPA) Working Paper, Stanford University. https://cepa.stanford.edu/sites/default/files/York%20%28March%202014a%29_0.pdf

Appendix 1- Memorandum of understanding

Agreement to participate in the Early Years Toolbox Pilot Study

Please sign both copies, retaining one and returning the second copy to Dr Clare Huxley at: Institute for Employment Studies, City Gate, 185 Dyke Road, Brighton BN3 1TL

Setting Name: _____

Aims of the pilot study

This project will train early years practitioners in the use of the Early Years Toolbox, a collection of apps designed to measure young children's development. By providing staff with information about children's emerging abilities, the idea is that practitioners will be better placed to ensure children's experiences in the settings are tailored to their individual development needs.

Funded by the Education Endowment Foundation (EEF), the key aims of this pilot study are to explore whether the programme shows evidence of promise, whether it is feasible for early years settings to deliver as part of routine practice, and whether such a programme could be implemented on a wider scale.

The Early Years Toolbox

The Early Years Toolbox comprises a suite of apps designed to measure young children's emerging cognitive, self-regulatory, language and social development. Each measure is a brief, engaging, game-like assessment that has been developed for the iPad. Developed by leading early years researchers, Professor Edward Melhuish (University of Oxford) and Dr Steven Howard (University of Wollongong, Australia), the Toolbox is designed to capture abilities that research has shown to be most predictive of later academic, social, emotional, cognitive and life outcomes.

What does the study involve?

Around 25 early years settings will participate in this pilot study.

Professor Edward Melhuish from the University of Oxford, in partnership with Action for Children, will train staff in participating early years settings in the use of the Toolbox. Training will take place in October 2018 and will comprise a one day training session (venue details to be confirmed separately). Staff will be trained in three of the six apps that form part of the Toolbox, but will have access to all six apps to use with children in their setting, if they wish to do so. Follow-up training will be provided as part of a visit to the setting by the Action for Children team.

Settings will be provided with an iPad and the necessary software for the app. The settings will be visited approximately every three months by a member of the Action for Children team, with further support provided through regular phone calls, plus access to a helpline via email and telephone. Nurseries will be expected to use the app with three year old children attending the setting, at least once per term.

The evaluation of the pilot is being conducted independently by a team of researchers from the National Institute of Economic and Social Research (NIESR) and the Institute for Employment Studies (IES). As part of the evaluation, nursery managers, teachers and staff working with three year olds will be asked to:

- Complete a short online survey prior to the initial training session, covering their experience as an early years practitioner
- Complete a short online survey towards the end of the programme, focusing on their experiences of the programme
- Participate in interviews with the evaluation team about their experiences of the programme.
  The evaluation team will carry out visits to some of the settings participating in the study, while other settings will be asked to participate in telephone interviews. Half of the visits and telephone interviews will take place in early 2019, with the remainder taking place towards the end of the programme in summer 2019.

A small number of settings may also be asked if a researcher from the evaluation team can be present to observe a quarterly visit from the project team.

Use of Data

All data collected during the study will be treated with the strictest confidence and in line with the Data Protection Act 2018 and the General Data Protection Regulation. The scores obtained on the games that form part of the Toolbox will be collected through the app and accessed by the project team (Professor Edward Melhuish and Action for Children) and the evaluation team (NIESR and IES). Information will also be collected on children's gender and date of birth. No individual setting, staff member or child will be identified in any report arising from the research. A letter explaining the study will be provided for settings to distribute to parents, this will explain that parents are able to withdraw their child's data from the study at any time and the process for doing so. Once the research is complete, all personal data collected as part of the study will be securely destroyed, within six months of completion of the study. The privacy notice for the study is available to download at: https://www.niesr.ac.uk/projects/pilot-evaluation-early-years-toolbox

Responsibilities

The project team will:

- Deliver training in the use of the Early Years Toolbox. This will comprise both an initial training day and follow-up training for participating staff
- Provide participating settings with an iPad and the required software
- Provide on-going support, including the provision of an email and telephone helpline; monthly phone calls, and quarterly visits
- Collect data from the Toolbox app from settings and share this with the evaluation team
- Be the first point of contact for any questions about the study
- Maintain a log of topics covered during quarterly visits/monthly calls

The evaluation team will:

- Provide an information letter to be sent to parents of children participating in the project
- Ensure all staff visiting settings are trained and have received DBS clearance
- Collect and analyse the data from the project and write up the findings
- Disseminate findings from the study

The setting will:

- Release staff so that they can attend the required training sessions, as well as providing staff with time to take part in the monthly calls and quarterly visits, and to meet the requirements of the evaluation. Action for Children will liaise with you to establish which staff will be attending the training, but we envisage that this will typically be the setting lead and the 3 year old room lead, or the class teacher and early years practitioner.
- Ensure staff who attend the training explain how to use the apps to other staff working with 3 year olds in the setting
- Ensure staff use the app with three year old children attending the setting, at least once per term during the period from October 2018-July 2019. Where feasible, we ask that you use the app with all three year old children attending the setting, but recognise that in some settings it will only be practical to use the app with one class or room. The Action for Children team will discuss this with you.
- Distribute information letters (provided by the evaluation team) to parents/guardians of all three year old children attending the setting
- Provide information on the number of three year old children attending the setting, and the number of children for whom permission to share data has been withdrawn by parents
- Ensure that no data on children for whom permission for data usage has been withdrawn is shared
- Liaise with the project and evaluation teams to find appropriate dates for visits and interviews
- Ensure the shared understanding and support of all setting staff for the project and personnel involved.
- Be a point of contact for parents / carers seeking more information about the project.

We commit to participating in the Early Years Toolbox pilot study as detailed above

Setting lead [NAME]: _____

Email address: _____

Phone number: _____

Signature:_____

Date: _____

Other relevant setting staff:

Name:

Job title/role: _____

Email address: _____

Signature:_____

Date:_____

Name:

Job title/role: _____

Email address: _____

Signature:_____

Date:_____

Appendix 2- Parental withdrawal letter

Dear Parent/Guardian,

We are writing to inform you about an exciting new project that your child's nursery is taking part in.

Over the next year, staff at the nursery will be trained in using the Early Years Toolbox. The Toolbox is a series of apps designed by leading early years researchers to help staff measure and understand young children's development. The app consists of a series of games which are designed to be engaging and fun for children. The aim is that by using the Toolbox to obtain information on children's development, staff will be better placed to ensure that children experience activities and opportunities that best meet their individual needs.

The Education Endowment Foundation (EEF) is funding a pilot study of this project to examine the use of the Toolbox in early years settings. The project is a collaboration between Professor Edward Melhuish from the University of Oxford, one of the developers of the Early Years Toolbox, and Action for Children. The programme is being independently evaluated by a team of researchers from the National Institute of Economic and Social Research (NIESR) and the Institute for Employment Studies (IES). The evaluation will collect information on whether the programme shows potential to improve outcomes for children and whether it is feasible for nurseries to deliver.

As part of the study, staff at the nursery will use the Early Years Toolbox apps with your child, to explore their development. The results, along with information on your child's gender and date of birth (but not their name or identity), will be shared with the project team (Professor Edward Melhuish and Action for Children) and the evaluation team (NIESR and IES). The data are anonymised so that it will not be possible to identify individual children. Children's responses and all other data collected as part of the project will be treated with the strictest confidence and in line with data protection regulations. We will not use your child's name or the name of the nursery in any report arising from the research. The privacy notice for the study is available at: https://www.niesr.ac.uk/projects/pilot-evaluation-early-years-toolbox

We expect that your child will enjoy being part of the programme and using the app. You may request that your child's data be withdrawn from the study at any time.

If you prefer for your child NOT to be part of this study, please inform the nursery manager or class teacher by completing the form below.

If you would like more information about this project, please contact me using the details given below. The evaluation has been approved by the Ethics Committee of the National Institute of Economic and Social Research. If you require clarification of the ethical approval or have any concerns during the course of the research, please contact Dr John Kirkland, NIESR's Chief Operating Officer (j.kirkland@niesr.ac.uk).

Yours sincerely,

Lucy Stokes
Co-Principal Investigator, NIESR

Email:l.stokes@niesr.ac.uk
Telephone: 020 7654 1928

Withdrawal Form

Please only return this form to the nursery manager or class teacher if you are NOT willing for your child's data to be used as part of the Early Years Toolbox pilot study.

I **DO NOT** give permission for my child to use the Early Years Toolbox as part of the pilot study.

Your Child's Name ……

Your Name ……..

Your Signature ……

Parent / Guardian (delete as appropriate)

Date ……

Appendix 3- Privacy notice

Pilot evaluation of Early Years Toolbox: Privacy Notice

Why are we collecting this data?

The National Institute of Economic and Social Research (NIESR), in partnership with the Institute for Employment Studies (IES), are collecting personal data to enable the pilot evaluation of the Early Years Toolbox. The Toolbox is an app designed to measure young children's emerging cognitive, self-regulatory, language and social development. The aim of the programme, delivered by Action for Children together with Professor Edward Melhuish of the University of Oxford, is that by providing staff in early years settings with information about children's emerging abilities, practitioners will be better placed to ensure children's experiences in the settings are tailored to their individual development needs.

The key aims of the pilot evaluation are to explore whether the programme shows evidence of promise, whether it is feasible for early years settings to deliver as part of routine practice, and whether such a programme could be implemented on a wider scale.

Who is this research project sponsored and funded by?

NIESR and IES are undertaking the independent evaluation of the pilot, funded by the Education Endowment Foundation (EEF). NIESR and IES are joint data controllers for this evaluation, alongside Action for Children and the University of Oxford, who will be delivering the programme in early years settings.

What is the legal basis for processing activities?

The legal basis for processing personal data is covered by:

GDPR Article 6 (1) (f) which states that 'processing is necessary for the purposes of the legitimate interests pursued by the controller or by a third party except where such interest are overridden by the interests or fundamental rights and freedoms of the data subject which require protection of the personal data'.

Our legitimate interest for processing the personal data collected in this study is to conduct the evaluation.

How will personal data be obtained?

Personal data will be collected in two ways: from information provided directly by participating early years settings, and through the Toolbox app.

What personal data is being collected by this project?

Personal data for the evaluation will include data about practitioners and children from participating early years settings as described below:

Action for Children will collect data (name, job title and contact details) about nominated staff at the participating settings, and share this with the evaluation team (NIESR and IES). These data will enable the evaluation team to liaise with staff at the settings about the evaluation and to facilitate two online surveys of participating practitioners.

Children's scores on the games they play as part of the Toolbox app, along with personal data about children (gender and date of birth) will be collected through the Toolbox app; the delivery team (Oxford and Action for Children) will share this data with the evaluation team.

These data will enable the evaluation team to undertake the research and analysis required for the pilot evaluation.

Who will the personal data be shared with?

Personal data will be shared with the evaluation team (NIESR and IES) and the delivery team (Action for Children and the University of Oxford).

Data collected through the Toolbox app will be shared with the evaluation team using an encrypted USB drive. All other data will be transferred via the IES' OwnCloud platform. Only authorised users can access the Own Cloud system through an assigned user name and password.

Is personal data being transferred outside of the European Economic Area (EEA)?

No - the data collected through the Toolbox app will be in transit between the UK where it is collected by practitioners using the apps and NIESR for data analysis via an Amazon server in the U.S.A which is owned by the University of Wollongong (Australia) but the data will not be accessed or manipulated outside of the EEA. The delivery team at Action for Children are able to access the data on this server; these data will then be shared with the evaluation team using an encrypted USB drive exchanged in person at the Action for Children London office.

How long will personal data be retained?

NIESR and IES will delete any personal data six months following the completion of the project.

Can I stop my personal data being used?

The evaluation team handles your personal data in accordance with the rights given to individuals under data protection legislation. If at any time you wish us to withdraw your data or correct errors in it, please contact Lucy Stokes (l.stokes@niesr.ac.uk)

In certain circumstances, data subjects have the right to restrict or object to processing, please contact NIESR's Data Protection Officer. They also have the right to see information held about them. The evaluation team will cooperate fully when a subject access request (SAR) is made.

Who can I contact about this project?

Please contact Lucy Stokes (l.stokes@niesr.ac.uk) for any queries relating to this project.

In certain circumstances, data subjects have the right to restrict or to object to data processing. Please contact NIESR's Data Protection Officer in these circumstances. Individuals also have the right to see information held about them. You can make a subject access request by contacting NIESR.

If you have a concern about the way this project processes personal data, we request that you raise your concern with us in the first instance (see the details above). Alternatively, you can contact the Information Commissioner's Office, the body responsible for enforcing data protection legislation in the UK, at https://ico.org.uk/concerns/.

This privacy notice was last updated on 10th April 2019.

Appendix 4 – Survey questionnaires

Practitioner survey - baseline

Your role and your organisation

1. What is your name?

2. What is the name of your setting?

3. What best describes your gender?
   a. Female
   b. Male
   c. Other
   d. Prefer not to say

4. What is your age?
   Under 18
   18 to 24
   25 to 29
   30 to 39
   40 to 49
   50 to 59
   60 and over
   Prefer not to say

5. What is your job role(s)? (Please select all that apply).
   a. Early Years Practitioner (Level 2)
   b. Early Years Practitioner (Level 3)
   c. Early Years Practitioner (Level 4 and above)
   d. Early Years Practitioner (unqualified/apprentice)
   e. Specialist role, e.g. Special Education Needs and Disability Lead (SEND, Safeguarding, Literacy or Maths Leader
   f. Deputy Manager
   g. Manager
   h. Early Years Teacher
   i. Teacher (QTS)
   j. Teaching assistant
   k. Other, please specify

6. Which is the highest level of Early Years qualification you hold?
   a. Level 2
   b. Level 3
   c. Level 4
   d. Level 5 diploma, Foundation degree
   e. BA, BSc
   f. Early Years Professional Status (EYPS), Early Years Teacher Status (EYTS), Qualified Teacher Status (QTS)
   g. Level 6 Postgraduate Certificate/Diploma, Professional Graduate Certificate
   h. Level 7 Postgraduate Certificate/Diploma
   i. Other Postgraduate qualification
   j. None, but I am working towards an Early Years qualification
   k. None, and I am not working towards an Early Years qualification
   l. Other, please specify

7. How many years have you been working as an Early Years professional?
   [Less than one year, one year 1 day to three years, three years 1 day to five years, five years one day to ten years, more than ten years]

Assessing children's development in the early years

8. How confident do you feel in your ability to assess children's development in the early years in the following areas:
[Very confident, quite confident, somewhat confident, not very confident, not at all confident, don't know]
   a. Language ability/ Vocabulary range
   b. Numeracy skills
   c. Self-regulation
   d. Social development

9. How confident do you feel in your ability to identify early years children who are most in need of support in the following areas: [Very confident, quite confident, somewhat confident, not very confident, not at all confident, don't know]
   a. Language ability/ Vocabulary range
   b. Numeracy skills
   c. Self-regulation
   d. Social development

10. Have you received any training on assessing children's development in the early years?
   a. Yes
   b. No
   c. Not sure

11. [If Q10 = a] Please could you briefly describe any training you have had
[open text]

12. Do you have experience of using apps to assess children's development in the early years as part of your work?
   a. Yes, I have used one or more apps for several years
   b. Yes, I have used one or more apps for around a year
   c. Yes, I have started using one or more apps this year
   d. No, I have never used an app for this
   e. No, but I have used an app for this as part of a research task.
   f. Not sure.

13. [If Q12 = a-c] Please give examples of apps you have used to assess children's development in the early years
[open text]

14. Do you have experience of assessing children's development in the early years using other methods, e.g. using a non app-based test, questionnaire or observation?
   a. Yes
   b. No
   c. Not sure

15. [If Q14 =a] Please give examples of non app-based methods you have used to assess children's development in the early years
[open text]

16. Please indicate how often, on average, you assess children's development in the early years in the following areas.
[More than once a month, Once a month, Every two months, Every term, Twice a year, Once a year, Less than once a year, not sure.]
   a. Language ability/vocabulary range.
   b. Numeracy skills
   c. Self-regulation
   d. Social development.

17. Please indicate how often you would like to assess children's development in the early years learners in the following areas.
[More than once a month, Once a month, Every two months, Every term, Twice a year, Once a year, Less than once a year, not sure.]
   a. Language ability/vocabulary range
   b. Numeracy skills
   c. Self-regulation

d. Social development

18. Have you faced any barriers to assessing children's development in the early years? Please tick all that apply.
    a. Time to complete the assessments
    b. Time to act on learning from the assessments
    c. Lack of management support to complete the assessments
    d. Anti- assessment attitudes from other staff at your setting
    e. Anti- assessment attitudes from parents of the early years learners
    f. Other, please specify

19. I would like more help in assessing children's development in the early years in the following areas:
    [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

Working with Early Years children

Please indicate to which extent you agree with the statements.

20. I give support to every early years child I am responsible for with:
    [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development.

21. I know where to get resources to help support early years children with:
    [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

22. I am aware of interventions to help support early years children with:
    [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

23. I would like more help in supporting early years children with:
    [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

24. [If Q23 =a,b,c have the response strongly agree or agree] What other support or training would you like to help you support early years children with language ability/vocabulary range, numeracy skills, self- regulation or social development?
    [open text]

Practitioner survey - Endpoint

Your role and your organisation

1. What is your name?

2. What is the name of your setting?

3. What best describes your gender?
   a. Female
   b. Male
   c. Other
   d. Prefer not to say

4. What is your age?
   Under 18
   18 to 24
   25 to 29
   30 to 39
   40 to 49
   50 to 59
   60 and over
   Prefer not to say

5. What is your job role(s)? (Please select all that apply).
   a. Early Years Practitioner (Level 2)
   b. Early Years Practitioner (Level 3)
   c. Early Years Practitioner (Level 4 and above)
   d. Early Years Practitioner (unqualified/apprentice)
   e. Specialist role, e.g. Special Education Needs and Disability Lead (SEND, Safeguarding, Literacy or Maths Leader
   f. Deputy Manager
   g. Manager
   h. Early Years Teacher
   i. Teacher (QTS)
   j. Teaching assistant
   k. Other, please specify

6. Which is the highest level of Early Years qualification you hold?
   a. Level 2
   b. Level 3
   c. Level 4
   d. Level 5 diploma, Foundation degree
   e. BA, BSc
   f. Early Years Professional Status (EYPS), Early Years Teacher Status (EYTS), Qualified Teacher Status (QTS)
   g. Level 6 Postgraduate Certificate/Diploma, Professional Graduate Certificate
   h. Level 7 Postgraduate Certificate/Diploma
   i. Other Postgraduate qualification
   j. None, but I am working towards an Early Years qualification
   k. None, and I am not working towards an Early Years qualification
   l. Other, please specify

7. How many years have you been working as an Early Years professional?
   [Less than one year, one year 1 day to three years, three years 1 day to five years, five years one day to ten years, more than ten years]

Assessing children's development in the early years

8. How confident do you feel in your ability to assess children's development in the early years in the following areas:
   [Very confident, quite confident, somewhat confident, not very confident, not at all confident, don't know]
   a. Language ability/ Vocabulary range
   b. Numeracy skills
   c. Self-regulation
   d. Social development

9. How confident do you feel in your ability to identify early years children who are most in need of support in the following areas: [Very confident, quite confident, somewhat confident, not very confident, not at all confident, don't know]

a. Language ability/ Vocabulary range
b. Numeracy skills
c. Self-regulation
d. Social development

10. Since your setting started taking part in the Early Years Toolbox project, please indicate how often, on average, you have assessed children's development using the Early Years Toolbox in the following areas. [More than once a month, Once a month, Every two months, Every term, Twice a year, Once a year, Less than once a year, not sure.]
    a. Language ability/vocabulary range.
    b. Numeracy skills
    c. Self-regulation
    d. Social development.

11. Is your setting using any other assessment methods in addition to the Early years Toolbox?, e.g. observations.
    a. Yes (please specify) [open text box]
    b. No
    c. Not sure
12. Have you faced any of the following barriers when using the Early Years Toolbox? Please tick all that apply.
    a. Time to complete the assessments
    b. Time to act on learning from the assessments
    c. Lack of management support to complete the assessments
    d. Anti- assessment attitudes from other staff at your setting
    e. Anti- assessment attitudes from parents of the early years learners
    f. Other, please specify

13. I would like more help in assessing children's development in the early years in the following areas: [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

Working with Early Years children

Please indicate to which extent you agree with the statements.

14. I give support to every early years child I am responsible for with: [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development.

15. I know where to get resources to help support early years children with: [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

16. I am aware of interventions to help support early years children with: [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

17. I would like more help in supporting early years children with: [Strongly agree, agree, neither agree nor disagree, disagree, strongly disagree, not sure]
    a. Language ability/vocabulary range
    b. Numeracy skills
    c. Self-regulation
    d. Social development

18. [If Q17 =a,b,c d have the response strongly agree or agree] What other support or training would you like to help you support early years children with language ability/vocabulary range, numeracy skills, self- regulation or social development?
[open text]

Learning to use the Early Years Toolbox

19. [Mandatory] Did you attend the initial Early Years Toolbox training with Professor Ted Melhuish and staff from Action for Children?
    a. Yes
    b. No
    c. Not sure

20. [If yes to Q19 attended AfC training] Please indicate the extent to which you agree with the following statements
(Strongly agree, agree, Neither agree nor disagree, disagree, Strongly disagree, not sure)
    a. The Early Years Toolbox training was clear and useful.
    b. The training gave me a good understanding of the aims of the Early Years Toolbox programme overall
    c. The training gave me a good understanding of how I could use the Early Years Toolbox in my setting.
    d. After the training, I felt able to use the Early Years Toolbox to assess children at my setting.
    e. After the training, I felt confident that I could train other staff at my setting on how to use the Early Years Toolbox.
    f. After the training, I felt able to understand scores data from the Early Years Toolbox app

21. [If yes to Q19 attended AfC training] Were you involved in training other staff at your setting in how to use the Early Years Toolbox?
    a. Yes
    b. No
    c. Not sure

22. [If yes to previous question [Q21 training other staff] training other staff] How did you find the process of training other staff at your setting to use the Early Years Toolbox?
a. Very easy/Quite easy/Neither easy nor difficult/Quite difficult/Very difficult

23. [If 'yes' to question (q21) training other staff] How much ongoing support did staff at your setting need after you had trained them on how to use the Early Years Toolbox?
    a. None, a little, some, a lot

24. [If 'no' or 'not sure' to Q19 attended AfC training] Did a colleague show you how to use the Early Years Toolbox?
    a. Yes/ No/ Not sure

25. [If yes to Q24 colleague trained] Please indicate to what extent you agree with the following statements. After being shown how to use the Early Years Toolbox by a colleague, I felt able to…
    a. use the Early Years Toolbox to assess children at my setting.
    b. Understand scores data from the Early Years Toolbox app

5. Using the Early Years Toolbox

26. Which of the Early Years Toolbox apps have you used?
Please tick all that apply
    a. EYT Early Numeracy (numeracy app)
    b. Expressive Vocabulary (literacy app)
    c. Child Self- Regulation and Behaviour Questionnaire (self-regulation/questionnaire app)

27. How many children have you used the app with since you started using the Early Years Toolbox at your setting? (please note, if you have used the Toolbox with the same child several times, this counts as one child)
Please indicate this for each app.

EYT Early Numeracy (numeracy app), Expressive Vocabulary (literacy app), Child Self- Regulation and Behaviour Questionnaire (self-regulation/questionnaire app)
    a. None
    b. Less than 5
    c. 5 to 9

d. 10 or more

28. Please indicate the extent to which you agree with the following statements (Strongly agree, agree, Neither agree nor disagree, disagree, Strongly disagree, not sure)
a. I have used information from the Early Years Toolbox to inform my planning around activities and topics.
b. I have used information from the Early Years Toolbox to inform how I support a particular child or children.
c. I have used information from the Early Years Toolbox to identify areas where I would like to increase my knowledge or experience as a practitioner

29. [If q5 = g, manager or f, deputy manager] Have there been any costs associated with the Early Years Toolbox and the project (beyond paying for the apps and attending the training)? If so, please detail them here.

[open text numeric currency]

30. [If q5 = g, manager or f, deputy manager] Approximately how much staff time has been spent on the Early Years Toolbox and associated activities this academic year (including attending training and/or taking part in any follow-up support visits as well as testing with children)?

Please briefly describe below.
[open text]

31. [If q5 = g, manager or f, deputy manager] How many staff are involved in the Early Years Toolbox project at your setting (including staff who are supporting the project as well as those who are using the Early Years Toolbox)?

Please briefly describe below.
[open text]

Your experience using the Early Years Toolbox

32. Please indicate the extent to which you agree with the following statements
(Strongly agree, agree, Neither agree nor disagree, disagree, Strongly disagree, not sure)
a. The Early Years Toolbox is easy to use.
b. Children find the EYT Early Numeracy (numeracy) app engaging and fun
c. Children find the Expressive Vocabulary (literacy) app engaging and fun
d. It is easy to fit Early Years Toolbox assessments into my usual teaching activities
e. I have found the scores data collected through the Early Years Toolbox app to be useful in my work.
f. I have found observations made while using the apps with a child to be useful in my work.

Your views on the Early Years Toolbox

33. Reflecting on your experience of using the Early Years Toolbox over the past school year, please indicate the extent to which you agree with the following statements
(Strongly agree, agree, Neither agree nor disagree, disagree, Strongly disagree, not sure)
a. Using the Early Years Toolbox has helped me improve my practice.
b. Using the Early Years Toolbox has helped improve practice for all staff working with this age group at my setting.
c. Using the Early Years Toolbox has had an overall positive impact on learning and/or behavioural outcomes among the children we have used it with at my setting.
d. Using the Early Years Toolbox has had an overall positive impact on learning and/or behavioural outcomes in cases where individual children needed more support.
e. Using the Early Years Toolbox has greater benefits for children than other assessment approaches.

34. Are there any other comments you would like to make regarding your experiences of the Early Years Toolbox?

[open text]

Appendix 5 – Analysis of EYT app data

In this appendix we summarise how the app data were used and analysed as part of the pilot.

Data from the apps were provided to the evaluation team at two points during the pilot. An interim version of the data was supplied in April 2019, so that the evaluation team could begin to setup the data for analysis. A final version of the data was supplied in August 2019, once the delivery of the pilot was complete. These data covered the full period of the intervention and it is this final version of the data used in all of the analysis presented in this report.

Data were provided for each of the three apps that formed part of the pilot. The first row of Table 5.1 shows the original number of observations for each of the three apps in the final data files provided. It is already apparent from these figures that the numeracy and vocabulary apps had been used more frequently than the self-regulation app.

Table 5.1: Summary of app data and analysis sample

	Numeracy	Vocabulary	Self-regulation
Total number of observations	1,180	1,204	875
N observations removed (indicated as test/unusable identifier)	*89*	*145*	*93*
Sample for analysis of app duration	1,091	1,059	782
N observations where setting identifier invalid	*41*	*55*	*25*
N observations missing date of birth	*153*	*159*	*118*
Sample for analysis of frequency of use	897	845	639

When using the apps, practitioners were asked to enter a unique identifier for each child. However, this was not always entered in the format requested. In some cases this identifier field indicated that the game was a test or trial run. In our analysis we exclude those observations where the identifier clearly indicated a test, or where there was no real attempt to follow the instructions given for the format of the identifier (which may have potentially also indicated that this was a practice run of using the app). A small number of additional observations were identified in the field for study name as a test run. The total number of observations excluded on this basis are shown in the second row of Table 5.1. The remaining sample is that used for our analysis of the time it took to complete the apps (note that as a robustness check, we also ran this analysis on the same sample used for the analysis of frequency of use, but this made no substantive difference to the results).

For our analysis of frequency of use of the apps, we undertake some further cleaning of the identifiers, as this analysis relies on being able to accurately identify the same children over time. A cleaner identifier was created by extracting the first two digits of the entered identifier, which are intended to represent the setting identifier, and combining this with information on child date of birth, in a consistent format, to create a new unique identifier for each child. By doing so, this aimed to reduce the number of instances where observations would not be identified as relating to the same child, purely, for example, due to differences in the format of the identifier entered over time.[12]

The first two digits of the individual identifiers were intended to be a setting identifier. However, on this basis, not all setting identifiers were in the expected range (further investigation of the data suggested this was often due to, for example, just a child's date of birth being entered in this field). The number of observations affected by this was relatively small (41 for the numeracy app, 55 for the vocabulary app and 25 for the self-regulation app; fourth row of Table 5.1).

[12] Although note that in a case where two children at the same setting had the same date of birth, this could overstate the number of times the app was used with an individual child, however, the number of such cases is expected to be very small.

Furthermore, date of birth was often missing from the app data. Once these observations were also excluded, this resulted in the following number of observations with a cleaned unique identifier: 897 for the numeracy app, 845 for the vocabulary app and 639 for self-regulation app (final row of Table 5.1). It is this sample that we use in our analysis to produce estimates of frequency of use. As a useable identifier could not be produced for all observations, we acknowledge that it is likely that estimates of the frequency of use of the apps will be an underestimate.

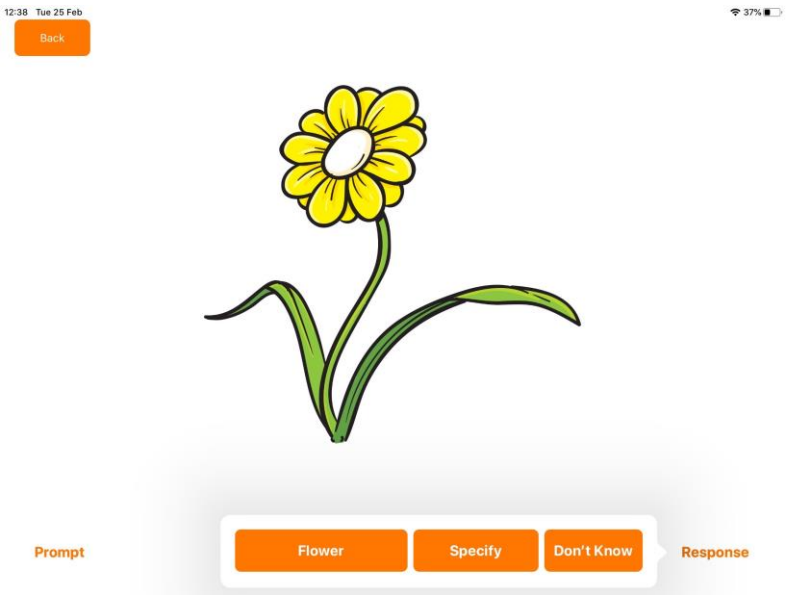
Issues for consideration for a future trial

The exploration of the app data has therefore highlighted some issues that would require consideration if a future evaluation or trial were to be undertaken. These are:

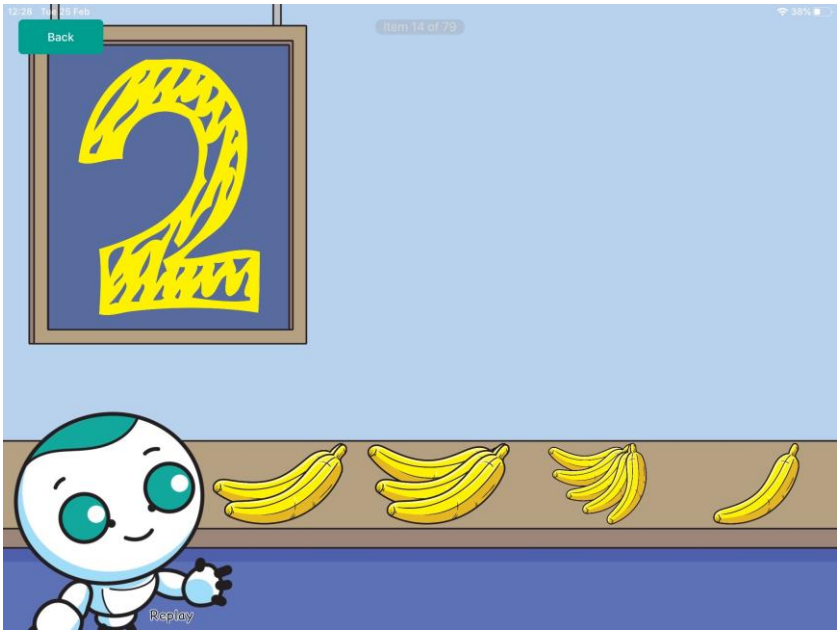
- **Setting and pupil identifiers**: it would be important to consider how ensure greater consistency in the way this information is entered into the app. For example, if it were possible to develop the app so that this field required a certain number of digits, or information to be entered in a particular format. Even more helpful, though potentially requiring considerable further app development, would be to enable practitioners to select a pupil from a drop-down list (however this could prove complex in practical terms due to the need to ensure names were only viewed by staff in the same setting). It would also be important to emphasise to practitioners the importance of entering this information consistently.
- **Missing information** (for example, date of birth): the figures presented above have shown that date of birth was missing for a significant proportion of observations. Again, the importance of entering this information could be emphasised to practitioners. In a trial, it is likely that information on date of birth would be collected separately as well through a separate data collection exercise as it would also be necessary to collect this information from a control group that were not using the apps. Moreover, the fact that information on date of birth was often missing is not only important from the perspective of being able to link data on the same children over time, but also because the results of the app need to be interpreted on the basis of children's age.

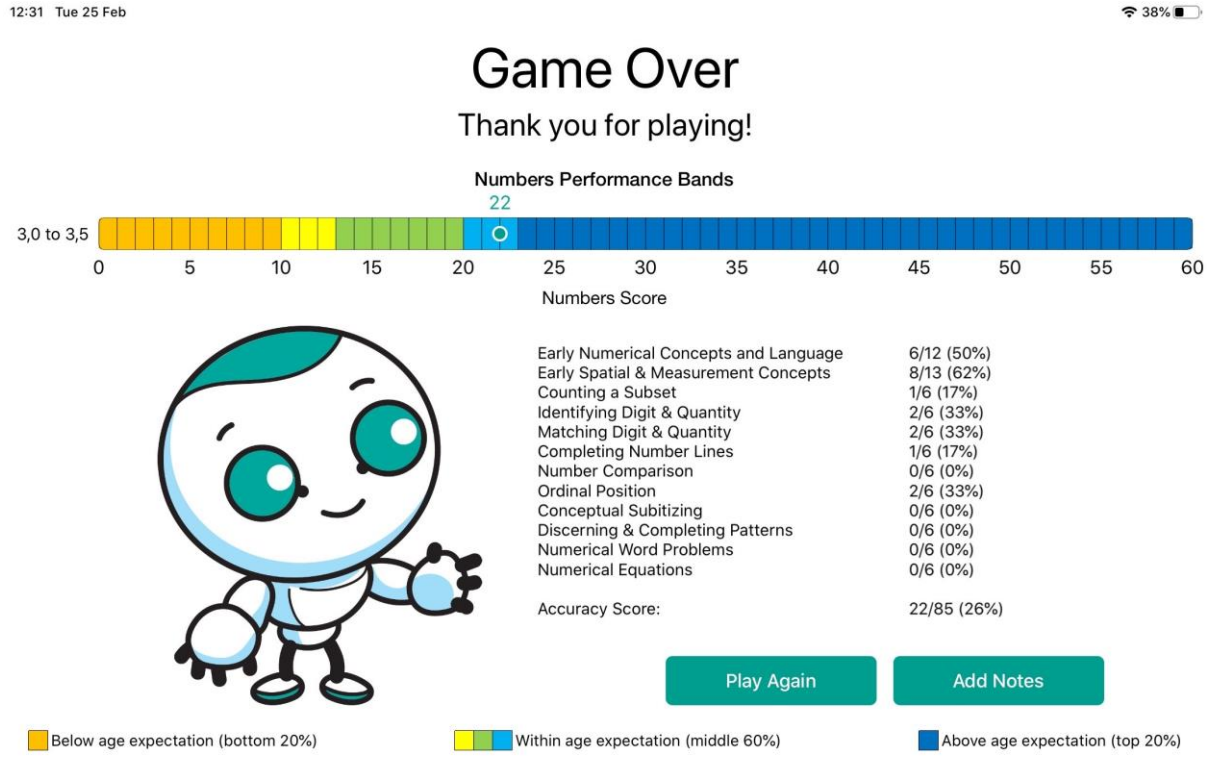
Appendix 6 – Example screenshots of the EYT

Vocabulary app



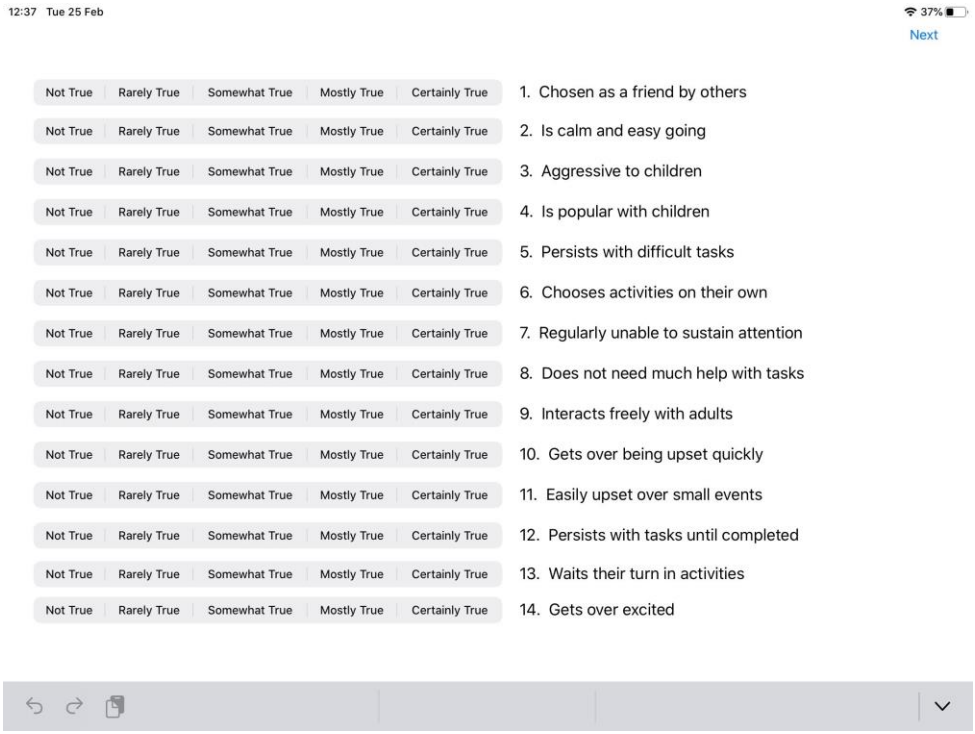
Numeracy app





Example output from the app

Self- regulation app



Example questions from the self- regulation app.

This document is available for download at **https://educationendowmentfoundation.org.uk**

The Education Endowment Foundation
5th Floor, Millbank Tower
21–24 Millbank
London
SW1P 4QP

https://educationendowmentfoundation.org.uk

@EducEndowFoundn

Facebook.com/EducEndowFoundn