# Improving Working Memory Plus Arithmetic Statistical Analysis Plan

Evaluator (institution): RAND Europe

Principal investigator(s): Elena Rosa Brown

Education Endowment Foundation

Template last updated: August 2019

| | |
|---|---|
| **PROJECT TITLE** | Improving Working Memory plus Arithmetic |
| **DEVELOPER (INSTITUTION)** | University of Oxford |
| **EVALUATOR (INSTITUTION)** | RAND Europe |
| **PRINCIPAL INVESTIGATOR(S)** | Elena Rosa Brown |
| **SAP AUTHOR(S)** | Dr Andreas Culora, Dr Sashka Dimova, Elena Rosa Brown, Merrilyn Groom |
| **TRIAL DESIGN** | Two-arm cluster randomised controlled trial with random allocation at school level |
| **TRIAL TYPE** | Effectiveness |
| **PUPIL AGE RANGE AND KEY STAGE** | 6-8, KS2 (Year 3) |
| **NUMBER OF SCHOOLS** | 201 |
| **NUMBER OF PUPILS** | 1,996 |
| **PRIMARY OUTCOME MEASURE AND SOURCE** | Number Skills (BAS3 number skills test) |
| **SECONDARY OUTCOME MEASURE AND SOURCE** | 1. Wider maths attainment (GL Assessment's Progress Test in Mathematics 8)<br>2. Working memory (Working Memory battery, central executive sub-texts)<br>3. Attention & behaviour (SNAP-IV Teacher Attention Rating Scale) |

## SAP version history

| VERSION | DATE | REASON FOR REVISION |
|---|---|---|
| 1.0 [*original*] | | *N/A* |

## Table of contents

SAP version history ..... 1
Table of contents ..... 2
Introduction ..... 3
Intervention ..... 3
Pupil selection ..... 4
Design overview ..... 5
Randomisation ..... 6
Outcome measures ..... 7
Baseline measures ..... 7
Primary outcome measure ..... 8
Secondary outcome measures ..... 9
Sample size and power calculations overview ..... 11
Analysis ..... 12
Primary outcome analysis ..... 12
Secondary outcome analysis ..... 13
Subgroup analyses ..... 14
Additional analyses ..... 14
Longitudinal follow-up analyses ..... 14
Balance at baseline ..... 14
Missing data ..... 17
Compliance ..... 18
Intra-cluster correlations (ICCs) ..... 19
Effect size calculation ..... 19
References ..... 20
Appendix A: Pilot testing of BAS3 ..... 21

# Improving Working Memory Plus Arithmetic Statistical Analysis Plan

Evaluator (institution): RAND Europe
Principal investigator(s): Elena Rosa Brown

Education Endowment Foundation

Template last updated: August 2019

## Introduction

The overall aim of this effectiveness trial is to assess whether the ‘Improving Working Memory plus Arithmetic’ (hereafter IWM+A) intervention leads to improvements in the number skills of Year 3 pupils, including specifically for pupils eligible for Free School Meals (hereafter FSM pupils). IWM+A is a multi-factorial regime intervention developed by researchers at Oxford University (Nunes et al. n.d.), targeting both Working Memory (hereafter WM) and arithmetic skills. Its aims are to improve the executive component of WM and to promote children’s skills in the organisation of information about numbers and about the operations of addition and subtraction in long-term memory.

A previous EEF efficacy trial was conducted of Improving Working Memory Plus (hereafter IWM+) (the same intervention as IWM+A, but with a different name) in which TAs who delivered the intervention were trained by Oxford University. The study found that pupils who received the intervention made three months additional progress in number skills compared to children in the control condition (d=0.24, -0.05, 0.52) (Wright et al., 2019).

Based on evidence from the efficacy trial, Oxford University was granted funding by the EEF to deliver IWM+A to more schools, using a train-the-trainer model, with Oxford University training Teacher Leaders (hereafter TLs) who in turn train Teaching Assistants (hereafter TAs) across 12 regions in England. This effectiveness trial builds on the IWM+ trial by ensuring the trial is powered to detect an effect on FSM pupils, as well as looking at how the intervention is delivered to more schools.

## Intervention

IWM+A is delivered in 10 one-to-one sessions, offered in addition to dedicated numeracy lesson time, with TAs working with two pupils over the course of an hour. TAs work directly with one child for half an hour while a second child plays targeted computer games in the same room. After half an hour the children switch activities. During the one-to-one element, TAs demonstrate and practise strategies with children, which are also reinforced through adaptive online games. The children access the targeted games online with an individualised login. The WM programme is individually paced, so that each child is automatically moved to the next level of the game when appropriate, in order to keep the level of challenge suitable for each child. The arithmetic games are also self-paced but are not computer-adaptive. Children receive a record sheet that is used to record the games they played; TAs guide them when they log in to the arithmetic games.

Children only move on to the arithmetic component (+A) after participating in the WM intervention. The minimum requirements to complete the programme is for pupils to receive all five WM combined with four of the five arithmetic sessions (nine sessions in total); though five WM combined with all five arithmetic sessions is considered optimal (10 sessions in total).

This effectiveness trial encompasses two phases: a train-the-trainer phase (Phase 1), and the full implementation phase (Phase 2). In Phase 1, Oxford University trained TLs (October 2020 – July 2021). TLs attended an initial session by Oxford University in May 2020 to support TLs approach and recruit schools to the trial. The session covered an overview of the programme and the design of the trial. TLs received recruitment fliers that they were able to adapt by including their contact details and then send to schools. This was meant to be part of the TLs’ training and to be followed by training sessions, but the process was interrupted after this initial session due to COVID-19.

The trainings were planned to be in-person sessions taking place in September 2020, but COVID-19 restrictions led the training to be delayed to February and April 2021, with the first two training sessions taking place remotely. TLs practised delivery in a small number of schools (not included as part of the Phase 2 of the trial) with three pairs of children per school: one pair in the morning, one pair after the morning break, and one pair in the afternoon. Oxford University were closely involved by undertaking observations to monitor implementation and undertake professional development conversations with each TL as part of their training in Phase 1. TLs received two additional training sessions in Phase 2 of the intervention (September 2021 – June 2022) to support their ability to deliver training, to monitor quality of the TA delivery, and to provide professional development to TAs. The two training sessions were convened immediately before TLs delivered training to TAs and Link Teachers - a teacher in the school responsible for supporting the TA.

As part of the trial, TLs train one Link Teacher and one TA nominated by each participating school allocated to the intervention group. Subsequently TAs deliver the intervention for one hour every week, over the course of 10 weeks between October 2021 and February 2022. The first five weeks of the intervention focus on improving WM (October-December 2021); the final five weeks focus on number and arithmetic operations (December 2021-February 2022).

In this effectiveness trial, IWM+A is being delivered to 100 schools, with another 101 schools assigned to the control group. The focus of this evaluation is to assess the effect of the intervention package by comparing pupils in schools receiving IWM+A to pupils in control schools receiving business as usual on the: (i) number skills of Year 3 pupils; (ii) maths attainment of Year 3 pupils; (iii) WM of Year 3 pupils; (iv) attention and behaviour among Year 3 pupils; (v) number skills of Year 3 FSM pupils.

The IWM+A intervention is led by Oxford University and is independently evaluated by RAND Europe. The study is funded by the Education Endowment Foundation (hereafter EEF).

For more information regarding the IWM+A intervention, study background and design, please refer to the IWM+A Evaluation Protocol (Brown et al., 2021).

### *Pupil selection*

Once schools had expressed interest in taking part in the trial and before randomisation, classroom teachers were asked to select children to participate in IWM+A. Oxford University provided schools with guidance on how to select appropriate children for the intervention. Teachers were asked to use their own judgement to nominate the 10 lowest-attaining students in mathematics at the end of KS1 (i.e., Year 2). The expectation was that teachers who were most familiar with pupils would be able to successfully select the most appropriate pupils with guidance; the delivery and evaluation teams did not perform any checks to verify if pupil selection was appropriate. In the previous efficacy trial teachers were asked to use KS1 assessments to select pupils, but these were cancelled in 2021 as a result of the impact of COVID-19, and therefore unavailable for use in this trial.

Pupils were selected in September 2021, at the start of the participating children’s Year 3, and prior to randomisation taking place. Schools were only able to nominate 10 children, regardless of the size of the school. For schools with more than one Year 3 form, teachers of all the classes were asked to cooperate and identify the 10 pupils showing lowest performance in mathematics across the year group.

In addition, the following exclusion criteria were observed:

- Deafness, blindness, and/or physical restrictions that might interfere with a child’s ability to use the online games.
- Behavioural problems that might interfere with a child’s ability to work independently and in the same room as another child.
- Children whose level of fluency in English would prevent them from engaging with the computer games.

All schools participating in the trial were asked to provide baseline data for the ten nominated pupils prior to randomisation. The extent of the baseline data required is outlined in the Outcome measures section below.

## Design overview

**Table 1. Trial design**

| | | |
|---|---|---|
| **Trial design, including number of arms** | | Two-arm cluster randomised controlled trial with random allocation at school level |
| **Unit of randomisation** | | School |
| **Stratification variable** | | **TL Region** (Cambridgeshire, Derbyshire, Devon, East Sussex, Essex, Greater Manchester, Merseyside, Nottinghamshire, South Yorkshire, Suffolk, West Sussex, West Yorkshire) |
| **Primary outcome** | variable | Number skills |
| | measure (instrument, scale, source) | British Ability Scales, Third edition (BAS3) (Number skills test, GL Assessment) |
| **Secondary outcome(s)** | variable(s) | 1. Wider maths attainment<br>2. Working memory<br>3. Attention and behaviour |
| | measure(s) (instrument, scale, source) | 1. Progress Test in Maths (PtM) (Level 8, GL Assessment)<br>2. Working Memory Battery (WMB) (Listening Recall, Counting Recall, Backward Digit Recall subtests, Pickering & Gathercole, 2001)<br>3. SNAP-IV Teacher Attention Rating Scale (adapted 15-item instrument[1] (Swanson et al., 2001; Bussing et al., 2008; Hall et al., 2020)) |
| **Baseline for primary outcome** | **variable** | Maths attainment |
| | measure (instrument, scale, source) | Early Years Foundation Stage Profile (EYFSP) (Numbers, and Shape, Space and Measures variables, Range 1-3, National Pupil Database) |
| **Baseline for secondary outcome** | **variable** | Attention and behaviour |
| | measure (instrument, scale, source) | SNAP-IV Teacher Attention Rating Scale (adapted 15-item instrument) |

This evaluation is designed as a TL region-stratified, two-arm, cluster randomised, effectiveness trial as outlined in Table 1. Random allocation was at the school level to avoid contamination, with schools randomly allocated to either the intervention group delivering IWM+A or to the control condition delivering business as usual. For more information on the randomisation approach and outcome, see the Randomisation section below.

This is an effectiveness trial, building on a previous EEF efficacy trial (Wright et al. 2019). Outcomes mirror those used in the efficacy trial, including number skills (primary outcome, measured using the

[1] We will use the same adapted scale that was used in the efficacy trial.

BAS3), WM (secondary outcome, measured using WMB) and attention and behaviour (secondary outcome, measured using SNAP-IV). Wider maths attainment will also be measured in this trial to understand how IWM+A affects maths attainment against the national curriculum (secondary outcome, measured using PtM8). More details on outcome measures are discussed in the Of the 201 schools that were randomised on 29th September 2021 as above, 195 nominated and provided baseline data for all ten selected pupils. Of the 6 schools not providing full data prior to randomisation, 4 schools had nominated and provided baseline data for nine out of ten pupils, while one school did so for six out of ten pupils and the final remaining school did so for four out of ten pupils. Thus, at baseline, the effective sample size was 1,996 pupils out of a theoretical population of 2,010 (201 schools multiplied by ten nominated pupils).

Outcome measures section below.

## Randomisation

Randomisation of schools to one of the two treatment arms took place on the 29th September 2021. In total, 201 schools were randomised to either the intervention or control group. Randomisation was conducted in Stata by a member of the evaluation team who was not blind to school identities, however, to verify independence, the randomisation code and outcome was checked by an independent staff member who was not part of the evaluation team. The randomisation was stratified by TL region with schools as the unit of randomisation, and (Year 3) pupils as the unit of analysis. A tailored package in Stata (`randtreat`) was used to implement the cluster randomisation with regional stratification. The code used to randomise schools as well as all relevant variables has been saved and can be made available if requested. This code will be included in the final Evaluation Report at the conclusion of the study. The allocation was recorded in Excel and communicated to the implementation team.

As specified in the Evaluation Protocol, randomising at the school level was perceived to reduce the chances of contamination between randomised groups as TAs would be delivering to all identified children in the school. Furthermore, regional stratification was deemed appropriate given the regional nature of delivery, with one TL delivering the training in each region (hence why we refer to 'TL regions'). Regions included in the randomisation as stratifiers were: Cambridgeshire, Derbyshire, Devon, East Sussex, Essex, Greater Manchester, Merseyside,[2] Nottinghamshire, South Yorkshire, Suffolk, West Sussex, West Yorkshire.

Randomisation occurred with a 50:50 allocation to treatment and control. Schools allocated to treatment receive IWM+A training and are expected to deliver the IWM+A intervention, while schools allocated to control are expected to carry on with business as usual (BAU), and do not receive the training. Schools assigned to the control group will be given £1,000 on completion of all endline assessments. These funds are to be used at the discretion of the school and could be used to buy an intervention programme of their choice, including IWM+A from September 2022. The value of the incentive is the same as was used in the efficacy trial.

Table 2 shows the actual allocations for the overall sample of participating schools and by TL region. In total, 101 schools were allocated to the control condition. A total of 100 schools were allocated to the intervention group. The numbers of schools in the trial varied by the stratification regions, between nine schools in Devon and 24 in South Yorkshire. The randomisation produced as equal an allocation to the intervention and control group as possible among the overall sample of participating schools and across the TL regions. For instance, in Devon, five of the nine schools in the stratum were allocated to the control group, and four were allocated to the Intervention group. In South Yorkshire, an equal number of schools (12) were allocated to each condition.

*Table 2. IWM+A randomisation results*

[2] Table 2 refers to two Merseyside TL regions as there are two TLs delivering the training in this area.

| | | Control group | Intervention (IWM+A) group | Total schools |
|---|---|---|---|---|
| 1 | Cambridgeshire | 9 | 9 | 18 |
| 2 | Derbyshire | 8 | 8 | 16 |
| 3 | Devon | 5 | 4 | 9 |
| 4 | Essex & Suffolk | 9 | 9 | 18 |
| 5 | Greater Manchester | 11 | 11 | 22 |
| 6 | Merseyside (LB) | 9 | 10 | 19 |
| 7 | Merseyside (LLS) | 9 | 9 | 18 |
| 8 | Nottinghamshire | 10 | 10 | 20 |
| 9 | South Yorkshire | 12 | 12 | 24 |
| 10 | Sussex | 9 | 8 | 17 |
| 11 | West Yorkshire | 10 | 10 | 20 |
| | Total | 101 | 100 | 201 |

Of the 201 schools that were randomised on 29th September 2021 as above, 195 nominated and provided baseline data for all ten selected pupils. Of the 6 schools not providing full data prior to randomisation, 4 schools had nominated and provided baseline data for nine out of ten pupils, while one school did so for six out of ten pupils and the final remaining school did so for four out of ten pupils. Thus, at baseline, the effective sample size was 1,996 pupils out of a theoretical population of 2,010 (201 schools multiplied by ten nominated pupils).

## Outcome measures

### *Baseline measures*

Baseline assessment for trial participating pupils *only* will consist of the following tests:

1) **Maths Attainment** based on the Early Years Foundation Stage Profile (EYFSP) Numbers, and Shape, Space and Measures subtests.

2) **Attention and Behaviour** based on the adapted 15-itme SNAP-IV Teacher Attention Rating Scale (Bussing et al., 2008; Swanson et al., 2001).

Baseline assessment was completed prior to randomisation: the Maths Attainment baseline – the EYFSP – was completed when children were in Reception as part of national testing requirements, whereas data on attention and behaviour was collected as part of this trial. The baseline measures are discussed in more detail below.

1) **Maths Attainment**

To reduce burden on schools it was decided to use existing Maths Attainment data available in the National Pupil Database (NPD). KS1 results would have been preferable given this would parallel the approach in the efficacy trial and the fact that they would be collected in the summer before implementation. However, in light of constraints resulting from the COVID-19 pandemic the 2021 KS1 SATs were cancelled by the Department for Education. Therefore, for this trial the Maths Attainment baseline data is sourced from the EYFSP completed in the 2018/2019 academic year, when the Year 3 children were in Reception. The EYFSP is a nationally administered measure of ability completed by practitioners (i.e. teachers or other qualified school staff) at the end of Reception when children are between the ages of four and five. The EYFSP is divided across the 17 early learning goals (ELGs) of the Early Years Foundation Stage (EYFS). For the purposes of this trial, we use two ELGs related to

Mathematics, namely: Numbers (MAT_G11), and Shape, Space, and Measures (MAT_G12). These scores are assigned by practitioners who are asked to decide whether children are meeting the expected learning levels. The NPD record for the Numbers and Shape, space and measures ELGs contain 1-3 assessment rating, where:

1. Indicates a child who is at the emerging level.
2. Indicates a child who is at the expected level.
3. Indicates a child who is at the exceeding level.

Children who have not been assessed due to long periods of absence or have been exempted are recorded as 'A' in the NPD. In this trial we will combine the score on both ELGs related to math to gain insight in the wider mathematical attainment. The total combined EYFSP score can therefore range from two to six.

We acknowledge that in using the EYFSP as baseline data, we will not be capturing current levels of mathematical attainment, particularly given the disruptions caused by school closures as a result of COVID-19. Furthermore, this is not a sensitive measure due to the restricted range of scores. If the EYFSP is not correlated with outcome measures (i.e., pre-post-test correlation is equal to 0), the study will still be powered to detect an effect of 0.196 on all pupils, and 0.249 on FSM pupils, as outlined in the Sample size and power calculations overview.

*2)* **Attention and Behaviour**

In this trial we use the SNAP-IV rating scale as a measure of attention and behaviour. The SNAP-IV is a rating scale developed in the United States, which aims to screen for hyperactivity, impulsiveness, and inattention (Swanson, 1992). It has a short and longer more comprehensive form and can be filled out by either the parent or teacher of children from the age of six and above. In this trial, we use the short version of the SNAP-IV completed by Teachers. The scale was also used in the previous efficacy trial.

Every item is organised as a four-point Likert scale ranging from 'not at all' to 'very much'. These responses are ranked from 0 to 3 with the total summed score varying from 0 to 45. All items receive equal weighting, in other words, the average rating per item based on the SNAP-IV is calculated by summing all items and dividing by the total number of items completed. Missing items in the SNAP-IV are excluded from the average rating per item. A higher score indicates that the child is at higher risk of hyperactivity or inattention.

Existing research indicates that the measure has good reliability across multiple samples (e.g., at .97 for the overall short scale) (Bussing et al., 2008). In this trial we focus on the items that measure inattention and hyperactivity, measured with a combined 18 items. As the scale was developed in the US, the evaluation team in the previous trial discarded three of these 18 items, which were judged to be irrelevant for English classrooms. We therefore use the same 15 item scale used in the previous efficacy trial.

Baseline data from the SNAP-IV Teacher Attention Rating Scale will be used as a baseline for the same measure at endline. The baseline testing using the SNAP-IV was completed in September 2021 by Teachers prior to randomisation. In the previous efficacy trial, the pre-test to post-test correlation was 0.683, but differed by treatment arm: for the control group, the SNAP-IV correlation was 0.74 for the control group, 0.67 for the Working Memory treatment and 0.66 for Working Memory and Arithmetic (Wright et al., 2019).

### *Primary outcome measure*

A post-test will be administered to assess pupils' number skills immediately after the end of the trial, in May 2022. For the purpose of this trial, the primary outcome will be assessed using the UK standardised number skills subscale of GL Assessment's British Ability Scales 3rd Edition (BAS3). The BAS3 number skills subscale was also used as the primary outcome in the efficacy trial, allowing for comparison across trials.

The BAS3 is a standardised battery of cognitive tasks that has long been established as a leading measure for assessing a child's cognitive ability and educational achievement (EEF, n.d.). The scales have been standardised on a sample of nearly 1,500 British children from diverse geographical areas and ethnic backgrounds (Swinson, 2013). Based on samples from standardisation of BAS2 an excellent overall reliability of 0.95 is reported for the number skills scale (Elliot & Smith, 2011b).

Two BAS3 versions exist. One version is suitable for the early years, and another for school-aged children (Swinson, 2013). In this trial we use the school-age version that is suitable for children from age five to 17 years and 11 months. Administration of the BAS3 number skills test takes less than ten minutes and will be undertaken on individual basis by trained professionals. Test administrators will be responsible for marking each question as the test is being administered (in line with the overall BAS3 testing protocol), ensuring that marking will also be blind to group allocation. More information on test administration is available in the trial protocol (Brown et al., 2021).

The BAS3 number skills achievement subscale is designed for adaptive testing: children start the test at a level considered appropriate for their age and are presented with easier items if they do not meet a criterion for moving to more difficult items; or are presented with progressively more difficult items until they can no longer meet the criterion for moving forward. The number skills scale has 51 items grouped into blocks of six items and one nine-item block. The suggested starting point for children in this trial given their ages (seven to eight years old) is item seven. Every child takes all items within a block and the next easiest block is administered until the child passes four or more items. Total scores are calculated by summing the number of questions answered correctly by pupils. Correct answers receive one point, while incorrect answers or no answers are scored as zero (Elliot & Smith, 2011). Given that this is a commercial test, it is not included in the appendix of the SAP.

Pilot testing was undertaken in Phase 1 to understand the extent to which scores of low attaining pupils were normally distributed. Details on this are presented in Appendix A.

Given the adaptive nature of the BAS3 number skills subscale, in this trial we use the age-standardised score in the primary outcome model.[3] This age-standardised score is a more useful measure than the raw or ability scores also available for the subscale, as the age-standardised score makes an allowance for the different ages of the pupils and can be compared against nationally representative samples, i.e., can easily identify if pupils score below or above the national average. The age standardised score for BAS3 ranges from 55 to 145.

### *Secondary outcome measures*

Three secondary outcomes will be captured as part of this trial in addition to the primary outcome:

1) **Wider maths attainment** based on the Progress Test in Maths (PtM) (Level 8, GL Assessments)

2) **Working memory** based on the Working Memory Battery (WMB) (for Children Listening Recall, Counting Recall, Backward Digit Recall subtests (Pickering & Gathercole, 2001))

3) **Attention and behaviour** as measured by the SNAP-IV Teacher Attention Rating Scale.

These are outlined further below.

#### 1) Wider maths attainment

[3] It is unclear which BAS3 score was used in the efficacy trial. Based on the pilot we have decided to use the age-standardised scores.

Although the intervention is designed to impact pupils' number skills rather than their wider mathematics attainment, in order to explore the impact of IWM+A on maths attainment in general, a wider measure of maths attainment will be used to analyse if IWM+A supports maths learning more generally. Wider maths attainment will be measured using GL Assessment's Progress Test in Mathematics (PTM). GL PTM was selected as the wider maths attainment measure because it is widely used by schools and it is well aligned to the national mathematics curriculum. PTM has been standardised on a sample of British children from diverse school level country, independent/grammar schools and previous attainment (Swinson, 2013). The measure adequately measures generic maths skills and reports for the same indicate good face validity, and excellent internal consistency.

In consultation with GL it was decided that Version 8, designed to be used by Year 3 children in the spring/summer term, would be used despite the lower ability level of the pupils. It was felt that Version 8 was the most appropriate and that the underlying constructs and the range of items available would mitigate against potential floor effects. The test takes approximately 60 minutes.

The group-delivered test will be administered to all ten selected pupils in each school at the same time under exam conditions by independent test administrators from AlphaPlus who will be blinded to allocation status. Tests will be marked by GL markers who will not know school allocation, thus adding an extra layer of blinding. EYFSP will be used as a baseline.

Despite the fact that the age-standardised score allows for benchmarking against national levels and trends, for this analysis the raw score will be used, to mitigate against prior instances of floor effects when used with lower-attaining pupils (Hodgen et al., 2019). Further details on the analysis plans are outlined in the Secondary Outcome analysis section.

### 2) Working memory

Pupils' WM will be measured using the WM Test Battery for Children (WMTB-C) (Pickering & Gathercole, 2001), using three sub-scales as per the efficacy trial: listening recall, counting recall and backwards digit recall. The WMTB-C was designed by Gathercole & Pickering (2000a). This is a standardised measure validated on a sample of 750 children from English schools (Alloway 2007). The test-retest reliability is high for listening recall (0.83) and counting recall (0.74), and lower for backwards digit recall (0.53). The WMTB-C is composed of nine subtests in total. The three subtests used in this trial measure WM directly (Gathercole & Pickering, 2000b).

The test will be administered at endline on a one-to-one basis by independent test administrators blinded to allocation status. Test administration takes around ten minutes. Given the high burden of test administration it was decided not to conduct the test at baseline. More information on the measure and its administration is available in the IWM+A Evaluation Protocol.

As in the efficacy trial (see Wright et al., 2019), to obtain a single measure of WM for the purposes of our analysis, we will employ confirmatory factor analysis (CFA) to derive a single, overall recall factor score using the total scores from each subtest. We hypothesise that performance in any sub-scale is driven by an underlying, general recall ability, and so expect all three sub-scales to load on to a single latent factor. As the WM test is adaptive, and subtests have different potential scores, using a total factor score avoids undue weighting of any content or mode of testing on the overall measure. We will use the recall factor derived using factor analysis if CFA confirms the factor structure, as measured using the Chi-square test, Root Mean Square Error of Approximation (RMSEA) and CFI (Comparative Fit Index). As a sensitivity analysis, we will replicate the analysis using the individual sub-scale scores in place of the single recall factor. The additional sensitivity analysis will allow us to examine to what extent any effects are unique to the empirical construct of a single, latent factor. CFA will be implemented using the `sem` or `gsem` command in Stata.

### *3) Attention and behaviour*

Attention and behaviour outcomes will be measured using the SNAP-IV Teacher Attention Rating Scale. As outlined in the theory of change (see Figure 1 in the IWM+A protocol), attention in class is considered to be an important proximal outcome leading to improved impact in maths. Further details on this measure can be found under Baseline measures.

The scale needs to be completed by an individual (i.e. a teacher) who is familiar with the pupil and their behaviour, and as such will be completed by pupils' teachers, who will consequently not be blind to allocation. Teachers will be asked to complete these measures for all nominated pupils in May 2022.

## Sample size and power calculations overview

The initial power calculations in the Protocol (Brown et al., 2021) were based on both the information provided in the Invitation to Tender and the subsequent set-up meetings with the Delivery team and the EEF. Power and minimum detectable effect size (MDES) calculations were performed using the PowerUp! Tool (Dong & Maynard, 2013).

At protocol stage, upper and lower bound MDESs were calculated using a two-level random assignment assuming equal allocation to the intervention and control groups. The upper bound MDES calculations were based on assumed sample sizes of 240 schools equally allocated across intervention and control, with 10 pupils entered into the trial per school. The lower bound MDES calculation assumed sample sizes of 200 schools again equally allocated across treatment and control, with 10 pupils entered into the trial per school.

Based on the efficacy trial, we assumed the proportion of variance in level 1 outcomes explained by level 1 covariates $R_1^2$ was 0.25, equating to a pre-post-test correlation of 0.5, as per Table 3 (Wright et al., 2019).[4] We used two-level clustered designs and based our calculations on an intra-cluster correlation (ICC) of 16 per cent (0.16, as per Table 3).

Assuming a desired power of 80 per cent, alpha of 5 per cent, a continuous, normally distributed outcome as well as taking all other assumptions outlined above and in Table 3, the overall upper bound MDES calculation at protocol stage was *d*=0.172, while the lower bound MDES calculation was *d*=0.188. Using the same parameters and assuming that on average 3 FSM-eligible pupils per school would be entered into the trial (a rate of 30% FSM-eligibility), the upper bound MDES at protocol stage was *d*=0.221 while the lower bound MDES was *d*=0.242.

*Table 3. Power Calculations*

| | | Protocol | | | | Randomisation | |
|---|---|---|---|---|---|---|---|
| | | UPPER | | LOWER | | OVERALL | FSM |
| | | OVERALL | FSM | OVERALL | FSM | | |
| Minimum Detectable Effect Size (MDES) | | 0.172 | 0.221 | 0.188 | 0.242 | 0.188 | 0.231 |
| Pre-test/ post-test correlations | level 1 (pupil) | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 | 0.50 |
| | level 2 (class) | NA | NA | NA | NA | NA | NA |
| | level 3 (school) | 0 | 0 | 0 | 0 | 0 | 0 |
| | level 2 (class) | NA | NA | NA | NA | NA | NA |

[4] It should be noted that the efficacy trial used KS1 arithmetic scores as baseline, as opposed to EYFSP suggested here. This was used as the assumption for the correlation in the power calculations as there is no currently published data on EYFSP and BAS3 correlations.

| | | Protocol | | | | Randomisation | |
|---|---|---|---|---|---|---|---|
| | | UPPER | | LOWER | | OVERALL | FSM |
| | | OVERALL | FSM | OVERALL | FSM | | |
| **Intracluster correlations (ICCs)** | level 3 (school) | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 | 0.16 |
| **Alpha** | | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 | 0.05 |
| **Power** | | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 | 0.8 |
| **One-sided or two-sided?** | | Two | Two | Two | Two | Two | Two |
| **Average cluster size** | | 10 | 3 | 10 | 3 | 10 | 4 |
| **Number of schools** | intervention | 120 | 120 | 100 | 100 | 100 | 94 |
| | control | 120 | 120 | 100 | 100 | 101 | 95 |
| | **total** | 240 | 240 | 200 | 200 | 201 | 189 |
| **Number of pupils** | intervention | 1,200 | 360 | 1,000 | 300 | 992 | 403 |
| | control | 1,200 | 360 | 1,000 | 300 | 1,004 | 401 |
| | **total** | 2,400 | 720 | 2,000 | 600 | 1,996 | 804 |

At randomisation stage, there were 201 schools and a total of 1,996 pupils with baseline data. This represents an average of 9.93 (10) pupils per school. Assuming the same parameters as at protocol stage and given the achieved overall sample, the MDES is *d*=0.188. As such, the MDES for the overall sample as reported in the SAP is the same as the lower bound MDES as reported at protocol stage (*d*=0.188).

As also shown in Table 3, the number of schools with at least one FSM-eligible pupil at randomisation stage was 189 out of the 201 schools, with a total of 804 FSM pupils in the sample almost equally balanced across the intervention and control group (403 and 401 FSM pupils respectively). This represents an average of 4.25 (4) FSM pupils per school. Assuming the same parameters as at protocol stage and given the achieved sample of FSM pupils, the estimated MDES is *d*=0.231. This is roughly at the mid-point between the lower bound MDES for the FSM sample as estimated (*d*=0.242) and the upper bound MDES estimate (*d*=0.221) at protocol stage. This is likely due to the slightly higher rate of FSM eligibility (40.3%) as compared to the estimation at protocol stage (30%), despite the overall sample size aligning more with the lower bound MDES scenario. As such, the study should be powered to detect any meaningful difference for the FSM subgroup.

We acknowledge that the pre-post-test correlation is likely to be an overestimate, however, as there are no published correlations between BAS3 and EYFSP we have used data from the previous trial. To provide a more conservative estimate we have also run the power calculations with the assumption that there is no correlation. If the EYFSP is not correlated with outcome measures (i.e., pre-post-test correlation is equal to 0), the study will still be powered to detect an effect of 0.196 on all pupils, and 0.249 on FSM pupils.

## Analysis

### *Primary outcome analysis*

**RQ1: What is the difference in number skills measured by the number skills subtest of the British Ability Scales (BAS3) of pupils in schools receiving IWM+A in comparison to those pupils in control schools receiving business as usual?**

To address RQ1 this evaluation will use the age-standardised scores on BAS3 as a primary outcome for number skills, with prior attainment being accounted for by EYFSP scores (see Of the 201 schools that were randomised on 29th September 2021 as above, 195 nominated and provided baseline data for all ten selected pupils. Of the 6 schools not providing full data prior to randomisation, 4 schools had nominated and provided baseline data for nine out of ten pupils, while one school did so for six out of ten pupils and the final remaining school did so for four out of ten pupils. Thus, at baseline, the effective sample size was 1,996 pupils out of a theoretical population of 2,010 (201 schools multiplied by ten nominated pupils).

Outcome measures section for more detail on the BAS3 and EYSFP). The pupil level BAS3 scores will be age-standardised with a mean of 100 and standard deviation of 15.

To address the primary research question an intention-to-treat (ITT) analysis will be undertaken using multi-level modelling (MLM) with fixed effects and random intercept. In ITT, data is analysed according to the group randomised, regardless of whether the treatment was actually received as intended, and irrespective of withdrawal from the intervention post-randomisation, or deviations in programme implementation. This principle is key in ensuring an unbiased analysis of intervention effects and is in line with the EEF's guidance (see EEF 2018).

To account for the nested nature of the data, a hierarchical linear model with two levels (school, pupil) will be fitted controlling for pre-intervention scores at the pupil level. Following the stratified randomisation, the model will include a term identifying the regions (strata).

Impact will be estimated by fitting the model in equation (1). Equation (1) is known as a 'random intercepts' model because $\beta_{0j}=\beta_0+u_j$ is interpreted as the school-specific intercept for school $j$ and $\beta_{0j}\sim_{i.i.d} N(\beta_{0,\sigma u}2)$ is random (it is a number that can take any value):

$$(1) \qquad Y_{ij} = \beta_0 + \mathrm{IWMA}_j\tau + Z_j\beta_1 + X_{ij}\beta_2 + u_j + e_{ij}$$

Where:

$Y_{ij}$= BAS3 scores on the number skills subtest for child $_i$ in school $_j$;

$\beta_0$ = the cluster-level coefficient for the slope of a predictor on number skills;

IWMA$_j$ = a binary indicator of the school assignment to intervention [1] or control [0];

$Z_j$ = school-level characteristics i.e. the stratifying variable of geographical location (as used for randomisation);

$X_{ij}$ = pupil level characteristics for pupil $_i$ in school $_j$, i.e. the pre-intervention EYFSP score;

$u_j$ = school-level residuals and

$e_{ij}$= individual-level residuals.

The coefficient τ in Equation 1 above will represent the outcome of the trial, with respect to the primary outcome measure. Equation 1 will also be replicated for each of the three secondary outcome measures (see next section).

The effect size (Hedge's *g*) will be calculated for τ. The effect size will be standardised using unconditional variance in the denominator and confidence intervals will be reported to communicate statistical uncertainty in line with EEF guidance (see EEF 2018). This will tell us the average effect of the intervention on pupil outcomes in treatment schools compared to those in control schools.

All analyses will be run in Stata, versions 17 onwards.

### Secondary outcome analysis

This effectiveness trial has three secondary outcome measures: wider math attainment; working memory; and attention and behaviour.

The secondary outcome analysis will follow the same procedures and use Equation (1) to estimate each respective secondary outcome model, as described in the primary analysis, substituting the primary outcome variable in turn with each of the secondary outcome variables.

For the secondary outcome analysis for Wider Maths and for Working Memory, the $X_{ij}$ vector for pupil $_i$ in school $_j$ will be represented by the EYFSP score as the baseline. Whilst neither Wider Maths scores nor Working Memory are age-standardised, in the main analysis we will not include age in the pupil-level characteristics, $X_{ij}$, as the trial is within one year group, somewhat minimising concerns over age effects. If we expect Wider Maths or Working Memory scores to be age-correlated, this may introduce bias into the measurement of $\tau$.[5] As an additional sensitivity analysis, we will run the regressions with age (in months) included as a covariate in $X_{ij}$We will also run a sensitivity analysis on the inclusion of EYFSP: given this was collected at the end of reception, it is possible that the correlation between EYFSP and current attainment is weaker than expected for a baseline measure, and may be introducing additional noise. For the secondary outcome analysis for Attention and Behaviour, the associated baseline measure will be used, since SNAP-IV has been administered at baseline.

### Subgroup analyses

As defined in the trial protocol, we will conduct a subgroup analysis for pupils eligible for FSM, using the same model as our primary analysis (Equation 1). With this analysis we will explore differential effects of IWM+A for FSM pupils as they are considered a key target group by the EEF. We will identify FSM pupils within the sample of intervention and control schools as they are registered in the NPD, using the variable EVERFSM_6_P.[6] Analysis will be undertaken using a binary FSM variable where FSM-eligible=1; non-FSM-eligible=0.

The study will first report mean outcomes by sub-categories of FSM-eligible children as a basic descriptive step. Then, we will enter an interaction term in Equation (1) above, to account for the FSM subgroup and treatment allocation while retaining the whole analytical sample in the model.

For the subgroup analysis, effect sizes and statistical uncertainty will be calculated and communicated as per the primary analysis (also see section on Effect size calculation).

### Additional analyses

In the protocol we suggested we would use factor analysis to understand the extent to which specific factors in PtM linked to IWM+A are impacted by the intervention. However, on reflection we believe the most appropriate and robust use of the data is to use same analytical model as the primary analysis. We believe this is correct for two reasons. Firstly, the BAS3 Number Skills subscale has been developed specifically to measure number skills and it is a reliable measure of this concept, in contrast our factor scores from the PtM would be exploratory as the PtM has not been used in this way widely. Secondly, interpretation of the same concept using two different measures feels unnecessary – especially given potential issues already discussed. For these reasons we think that we should prioritise the BAS3 as our primary outcome measure and use the PtM as a secondary outcome measure focusing on wider maths attainment.

[5] Due to randomisation, bias will only be introduced in $\tau$ if age is correlated with effectiveness of the IWM+A treatment (that is if older pupils improve more rapidly in response to the intervention). Whilst generally any omitted variable would introduce bias in all coefficient estimates, bias will not be introduced if the included variable is both uncorrelated with the omitted variable and all other covariates. By virtue of randomisation, treatment assignment should not be correlated with either school-level characteristics, $Z_{ij}$, or pupil-level characteristics (including age), $X_{ij}$. Thus the only reason $\tau$ would be biased would be if the treatment effect is correlated with age.

[6] NPD Alias for this variable is: EVERFSM_6_P_[term][yy]. The variable indicates whether the pupil 'has ever been recorded as eligible for free school meals on Census day in any termly or annual Census (including Alternative Provision Census and PRU Census where available) in the last 6 years up to the pupil's current year (not including nursery)'.

The protocol also outlines an additional compliance analysis. This is outlined below (see Compliance).

### Longitudinal follow-up analyses

Longitudinal analysis will be commissioned separately to this trial.

### Balance at baseline

A well-conducted randomisation should create groups that are equivalent at baseline, with any imbalance at baseline occurring by chance (Glennerster & Takavarasha, 2013). However, to check for, and monitor, imbalance at baseline in the realised randomisation, baseline equivalence testing will be conducted at the school and pupil level.

At the school level, we will check the balance in the following variables by means of cross-tabulations and histograms that assess the distribution of each characteristic between the control and intervention groups. We will also report associated effect sizes to quantify the magnitude of any observed imbalance (for continuous measures), but will not undertake statistical testing as the initial randomisation necessarily means that any imbalance we observe is due to chance and therefore statistical tests are not appropriate:

- Ofsted ratings;
- Proportion of children eligible for FSM.[7]

At the pupil level, balance will be assessed as above but for the following characteristics:

- Gender;
- Age (in months);
- Prior attainment using the EYFSP;
- SNAP-IV Teacher Attention Rating Scale;
- FSM status for pupils.

At the time of drafting this SAP, data on both of the school level characteristics and three of the five pupil level characteristics (gender; age; SNAP-IV Teacher Attention Rating Scale) listed above are available for preliminary analysis of balance at baseline.

[7] Variable used is PNUMFSMEVER: ‘Percentage of pupils eligible for FSM at any time during the past 6 years’.

Table 4 documents the balance between schools and pupils in schools allocated to the control and intervention conditions respectively.

Looking at the two school level characteristics first, Table 4 presents the balance across the intervention and control group with regards to Ofsted ratings and proportion of pupils eligible for FSM. With regards to Ofsted ratings, the proportion of intervention and control schools rated ‘Outstanding’ (14.3% and 15.0% respectively), ‘Good’ (67.3% and 71.0% respectively), ‘Requires improvement’ (13.3% and 11.0% respectively) and ‘Inadequate’ (5.1% and 3.0% respectively) were similar. Indeed, the largest percentage point difference was observed among schools rated ‘Good’, but this difference was still quite small at 3.7 percentage points.

Table 4 shows that the proportion of FSM-eligible children at the school level was similar in the intervention and control school (28.0% and 25.2% respectively) – a 2.8 percentage point difference was observed here. The assessment of these characteristics therefore suggests that, at the school level, a balanced distribution across the intervention and control groups was observed following randomisation.

The pupil level baseline characteristics also displayed a balanced distribution across the intervention and control groups. Along the lines of gender, the proportion of female and male pupils were similar in the intervention (53.0% and 47.0% respectively) and control groups (53.3% and 46.7% respectively). The distribution of pupil age in months was balanced across the intervention and control groups also; the mean age in months in the intervention group was 90.2 while this was 90.4 in the control group; or roughly 7.5 years in both groups. Finally, attention as measured by the SNAP-IV[8] was also balanced across the intervention and control groups. The mean on the SNAP-IV was 1.14 in both groups, indicating relatively low levels of attentional difficulty on average in both groups. The absence of any meaningful imbalance at baseline in terms of attention was highlighted by the effect size estimate (*g*=0.00) here, as shown in

[8] SNAP-IV attention ratings are as follows: 0=Not at all; 1=A little; 2=Mostly; 3=Very much. A higher score denotes greater attentional difficulty. The analysis of baseline equivalence was undertaken on continuous data as overall scores were used, derived by taking the mean across the 15-item attention scale.

Table 4.

We conclude that randomisation has resulted in intervention and control groups that are balanced and comparable based on the school and pupil level baseline characteristics available at the time of SAP drafting.

*Table 4. Baseline characteristics of groups as randomised*

| School-level (categorical) | National-level mean | Control group | | Intervention (IWM+A) group | | |
|---|---|---|---|---|---|---|
| | | n/N (missing) | Count (%) | n/N (missing) | Count (%) | |
| **Ofsted rating** | | | | | | |
| 1. Outstanding | 10.5 | 15/100 (1) | 15.0 | 14/98 (2) | 14.3 | |
| 2. Good | 65.8 | 71/100 (1) | 71.0 | 66/98 (2) | 67.3 | |
| 3. Requires improvement | 16.4 | 11/100 (1) | 11.0 | 13/98 (2) | 13.3 | |
| 4. Inadequate | 7.4 | 3/100 (1) | 3.0 | 5/98 (2) | 5.1 | |
| **School-level (continuous)** | | **n/N (missing)** | **Mean (SD)** | **n/N (missing)** | **Mean (SD)** | |
| **Proportion of FSM-eligible children** | 25.0 | 101/101 (0) | 25.2% (14.7) | 100/100 (0) | 28.0% (15.3) | |
| **Pupil-level (categorical)** | | **n/N (missing)** | **Count (%)** | **n/N (missing)** | **Count (%)** | |
| **Gender** | | | | | | |
| 1. Female | | 535/1,004 (0) | 53.3 | 526/992 (0) | 53.0 | |
| 2. Male | | 469/1,004 (0) | 46.7 | 466/992 (0) | 47.0 | |
| **Pupil-level (continuous)** | | **n/N (missing)** | **Mean (SD)** | **n/N (missing)** | **Mean (SD)** | **Effect size [95% CI]** |
| **Age (in months)** | | 1,004/1,004 (0) | 90.4 (3.8) | 990/992 (2) | 90.2 (3.7) | N/A |
| **SNAP-IV Teacher Attention Rating Scale** | | 1,004/1,004 (0) | 1.14 (0.75) | 992/992 (0) | 1.14 (0.77) | 0.00 [-0.09 – 0.08] |

As data becomes available[9] on the other characteristics of interest at baseline (additional to above table), we will assess baseline equivalence at the school level and the pupil characteristics as defined above.

In the final report, as above, we will not carry out statistical significance tests to assess balance at baseline, as the premise of statistical testing at baseline does not hold in randomised controlled trials.[10] Instead, a table of the means along with distributions (for continuous variables) or counts with

[9] EYFSP rating and FSM status will be obtained from the NPD.

[10] Baseline Data. *Consort (2010)*. Retrieved 26 March (2019): http://www.consort-statement.org/checklists/view/32-consort/510-baseline-data

percentages (for categorical variables) will be presented, as above (Senn, 1994).[11] While EEF guidance for assessing balance at baseline is such that differences in pupil-level pre-tests should be reported as effect sizes, the pre-test measure in this trial that will be used here (the EYFSP) is categorical, thus Hedge's g is not appropriate as it relies on data being normally distributed.

### *Missing data*

Missing data can arise from item non-response to any of the primary or secondary outcome measures, or from attrition of participants at school and pupil levels. Even though it is important to include all data, it can be problematic to apply the ITT principle if we are not able to complete follow-up testing for all randomised schools. To better understand the pattern of missing data and its impact on the analysis, we will explore the extent of missingness, and whether there is a pattern in missingness. We will analyse and report missingness for the primary outcome measure (and associated primary analysis); for the secondary outcome measures and the FSM subgroup we will report the extent of missing data using cross-tabulations but not undertake any additional analyses.

As a basic first step in the missing data analysis we will explore attrition across trial arms to assess bias (Higgins et al., 2011). We will provide cross-tabulations of the proportions of missing values on all baseline characteristics (as detailed in the previous section, at both pupil and school level), as well as on the primary outcome measure and secondary outcome measures.

To assess whether there are systematic differences between those who drop out and those who do not – and thus whether these factors should be included in analysis – we will model missingness at follow-up (defined as pupils with missing primary outcome data at endline) as a function of baseline covariates, including treatment. The analysis model for this approach will mirror the multilevel level model specified in Equation (1), with pupils clustered in classes, but the outcome will be a binary variable identifying missingness (where 1=missing; 0=complete) in a multilevel mixed-effects logistic regression model (using Stata's `melogit` command). We can similarly model missingness at baseline (defined as pupils with missing baseline data, namely missing EYFSP) as a function of endline covariates, including treatment.

We will follow the protocol for missing data suggested by the EEF (see EEF, 2018). For less than 5% missingness overall from randomisation to final analysis, a complete-case analysis will be employed. This assumes that data are missing completely at random (MCAR), which we will be able to test partially with the model outlined above. For more than 5% missing data overall from randomisation to final analysis, our approach will depend on pattern of missingness. If the pattern of missingness may be unrelated to the treatment effect (e.g., pupil illness, staff changes or COVID-related impacts on testing but not on intervention), then missing data will be assumed MCAR and we will continue with a complete-case analysis. Otherwise, we will adopt a full-information maximum likelihood (FIML) approach. FIML estimates the value of the trial parameter of interest ($\tau$ in Equation 1) by maximising the likelihood function in the presence of missing data. We will not supplement this analysis with a separate approach, usually multiple imputation, MI, analysis, as the results of the two have been shown recently to be broadly equivalent (Lee & Shi, 2021). Additionally, in comparison to MI, FIML produces the same results every time, can be estimated in a single model and simulation studies show that it can reduce bias as well as MI (for a detailed discussion of FIML vs MI see Allison, 2012 or Jakobsen et al, 2017). The FIML approach only alleviates bias if pattern of missingness is dependent only on observed variables, namely missing at random (MAR); if the reason for missingness is due to unobserved variables, namely missing not at random (MNAR), then FIML or MI will not improve estimates. Note, however, that MAR and MNAR are not distinguishable based on observed data. If it seems likely data could be MNAR, sensitivity analysis will be conducted and reported alongside headline estimates.

[11] There is a convention in some disciplines that a 10pp (or larger) difference in treatment and control means at baseline constitutes 'imbalance' is thus justification for including those measures in sensitivity analyses, but there are counter-arguments to this idea. See Roberts, C. and Torgerson, D. (1999) 'Baseline imbalance in randomised controlled trials', *BMJ*, 319:185; de Boer et al. (2015) 'Testing for baseline differences in randomized controlled trials: an unhealthy research behavior that is hard to eradicate', *International Journal of Behavioral Nutrition and Physical Activity*, 12:4.

The FIML approach will be implemented using the `sem` command in Stata.[12]

## Compliance

As the ITT approach captures the averaged effect of *offering* the intervention, we also propose to look at treatment effects in the presence of compliance at the pupil level to capture the average effect of *participation* in IWM+A.

For the purpose of this evaluation, we are employing the EEF's definition of compliance as 'the extent to which the critical ingredients of the intervention are delivered to and/or received by the target participants' (EEF, 2019b). In collaboration with the EEF and the delivery team at University of Oxford, we have defined 'compliance' as pupil attendance at sessions, which will be collected by TAs using attendance logs using a template developed by the delivery team. TAs will be asked to share the logs with the evaluator after all sessions have been completed. As shown in Table 5, children who have attended all five WM sessions and at least four arithmetic sessions will be marked as compliant; all those that do not obtain this will be marked as non-compliant. This is in line with the expectations set-out by the project team.

*Table 5. Pupil level compliance measure*

| Compliance criterion | Data source | Compliance indicator |
|---|---|---|
| **Attendance at WM sessions** | Register of pupil attendance recorded by TAs | Pupil has attended all five WM sessions |
| **Attendance at arithmetic sessions** | Register of pupil attendance recorded by TAs | Pupil has attended at least four out of five arithmetic sessions |

There will be two strands to the compliance analysis based on the measure outlined in Table 5. The core component of the compliance analysis will take compliance as a binary measure (where 1=compliant; 0=non-compliant, as defined Table 5), defined at the pupil level, based on attendance as recorded by the TAs. In a situation of imperfect compliance, whereby not all participating pupils are deemed compliant using the criteria outlined in Table 5, we will undertake a complier average causal effect (CACE) analysis, by drawing on an instrumental variable (IV) approach, and using a two-stage least squares (2SLS) estimation approach to recover the treatment effect for those who complied with assignment. The first stage of this approach estimates the extent to which the assignment to IWM+A encourages pupils to take up the treatment (the first stage regresses treatment assignment on compliance). This will estimate a compliance rate. Results for the first stage will report the correlation between the instrument and the endogenous variable; and an F test. The second stage of the IV estimation predicts the outcome as per Equation (1), but substitutes the treatment indicator ($\mathrm{IWMA}_j\tau$ in Equation (1)) with the compliance rate estimated in the first regression (Angrist & Krueger, 1991; Angrist, 2006). The results of this model will answer the research question: To what extent does *compliance with* IWM+A implementation requirements lead to improved outcomes for pupils? This model will be estimated for the primary outcome measure only.

We are also interested in understanding the role that 'dosage' (i.e. the amount of the intervention received by pupils) has on outcomes, given the important role programme participation can have on variance in outcomes, particularly for younger children (Zhi et al., 2010). If, as measured by the binary compliance indicator described above, the non-compliance rate exceeds 10%, we therefore propose as a second additional component to the compliance analysis to include data from pupil attendance records as a more detailed categorical variable when analysing the impact of IWMA+ on the primary outcome. This variable would take on one of eleven potential categories, from 0 (indicating the pupil has not attended any IWMA+ sessions) to 10 (indicating that the pupil has attended all five WM sessions and all five arithmetic sessions) with one-unit increments. The same analytical approach will be implemented using this more detailed dosage indicator as outlined for the core compliance model above (i.e. CACE analysis drawing on an IV approach and using 2SLS estimation). This will be an additional

[12] For more information on implementing missing data analysis using the FIML approach in Stata with the `sem` command, see here https://www.stata.com/meeting/switzerland16/slides/medeiros-switzerland16.pdf

analysis that will sit alongside the main component of the compliance analysis to better understand how different levels of session attendance affects outcomes. As with the core compliance model, this model will be estimated for the primary outcome measure only.

The limitation to estimating a dosage-varying local average treatment effect is that no differentiation is made between compliance with working memory sessions and compliance with arithmetic sessions. That is, a student who attends five working memory sessions and no arithmetic sessions would be treated identically to a student who attends five arithmetic sessions and no working memory sessions. Whilst we recognise this limitation, we believe dosage-varying effects may be justified for exploratory analysis under a situation where there is a large degree of non-compliance (over 10%), as it seems likely that any one-on-one time with a TA, on either the arithmetic or working memory component, could improve the primary outcome and thus be partially substitutable.

### *Intra-cluster correlations (ICCs)*

The ICC is a key parameter for clustered trials. It represents the proportion of variance in a given outcome that can be explained by the variation between clusters (i.e. schools) as opposed to within-clusters.

The ICC used for the power calculations reported in the section on Sample size and power calculations overview and at protocol stage (see the IWM+A evaluation protocol for more information) is based on the ICC reported at analysis stage in the previous EEF IWM efficacy trial, at 0.16.

In the final report we will report ICCs as at protocol stage; as at randomisation; and at analysis stage. The ICC at analysis stage will be based on the primary outcome measure; and will be calculated using: (i) the same model as Equation (1) and; (ii) a model similar to that documented in Equation (1) but with no covariates, accounting for the clustering of pupils in schools (the so-called empty model). ICCs will be estimated using Stata's `estat icc` command.

## *Effect size calculation*

We will use the effect sizes (hereafter ES) for cluster-randomised trials given in the EEF evaluator guidance, as adapted from Hedges (2007):

$$ES = \frac{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$$

Where $\boldsymbol{(\bar{Y}_T - \bar{Y}_C)_{adjusted}}$ is the mean difference between the intervention and control group adjusted for baseline characteristics and $\boldsymbol{\sqrt{\sigma_S^2 + \sigma_{error}^2}}$ is an estimate of the population standard deviation (variance).

From the primary outcome model, we will take each group's adjusted mean and variance to calculate the effect size. This variance will be the total variance (across both pupil and school levels, without any covariates, as emerging from a 'null' or 'empty' multi-level model with no predictors). The ES therefore represents the proportion of the population standard deviation attributable to the intervention (Hutchinson & Styles, 2010). A 95% CI for the ES, that takes into account the clustering of pupils in schools, will also be reported. Effect sizes will be calculated for each of the models estimated.

All ES will be estimated using the `eefanalytics` Stata package.[13]

[13] See here for more information https://ideas.repec.org/c/boc/bocode/s458904.html

# References

Allison, P. D. (2012). Handling missing data by maximum likelihood. *SAS global forum*, 2012 (312), 1038-21.

Angrist, J. D. (2006). Instrumental variables methods in experimental criminological research: what, why and how. *Journal of Experimental Criminology*, 2(1), 23-44.

Angrist, J. D., & Krueger, A. B. (1991). Does compulsory school attendance affect schooling and earnings?. *The Quarterly Journal of Economics*, 106(4), 979-1014.

Brown, E., Flemons, L., Leenders, E., Larmour, S., de Silva, A. (2021). Improving Working Memory plus Arithmetic Evaluation Protocol. London: *Education Endowment Foundation*. https://d2tic4wvo1iusb.cloudfront.net/documents/pages/projects/IWMA-protocol-final.pdf

Bussing, R., Fernandez, M., Harwood, M., Hou, W., Garvan, C. W., Eyberg, S. M., & Swanson, J. M. (2008). Parent and teacher SNAP-IV ratings of attention deficit hyperactivity disorder symptoms: psychometric properties and normative ratings from a school district sample. *Assessment*, 15(3), 317-328.

Dong, N., & Maynard, R. (2013). PowerUp!: A tool for calculating minimum detectable effect sizes and minimum required sample sizes for experimental and quasi-experimental design studies. *Journal of Research on Educational Effectiveness*, 6(1), 24-67.

Elliott, C. D., & Smith, P. (2011a). *British Ability Scales 3: Administration and scoring manual*: GL Assessment.

Elliott, C. D., & Smith, P. (2011b). *British Ability Scales 3: Technical Manual*: GL Assessment.

Glennerster, R. & Takavarasha, K. (2013) *Running Randomized Evaluations: A Practical Guide*. London: Princeton University Press.

Jakobsen, J. C., Gluud, C., Wetterslev, J., & Winkel, P. (2017). When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts. *BMC medical research methodology*, 17(1), 1-10.

Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32, 4:. 341 - 370 https://doi.org/10.3102/1076998606298043.

Hutchison, D., & Styles, B. (2010). *A guide to running randomised controlled trials for educational researchers*. Slough: NFER.

Lee, T., & Shi, D. (2021). A comparison of full information maximum likelihood and multiple imputation in structural equation modelling with missing data. Psychological Methods, 26(4), 466 – 485.

Nunes, T. et al. (n.d.). *Improving Working Memory and Arithmetic Knowledge*. Oxford: University of Oxford.

Senn, S. (1994). *Testing for baseline balance in clinical trials*, Statistics in Medicine, 13: 1715-1726.

Swanson, J. et al. (2001). Clinical Relevance of the Primary Findings of the MTA: Success Rates Based on the Severity of ADHD and ODD Symptoms at the End of Treatment. *Journal of the American Academy of Child & Adolescent Psychiatry*, 40(2), 168-79.

Swinson, J. (2013). British Ability Scales 3. *Educational Psychology in Practice*, 29(4), 434-435.

Wright, H. et al. (2019). *Improving Working Memory: Evaluation Report*. Education Endowment Foundation. Retrieved 29 November 2021, from: https://educationendowmentfoundation.org.uk/public/files/Projects/Evaluation_Reports/Improving_Working_Memory_Report_final.pdf.

# Appendix A: Pilot testing of BAS3

To understand better the psychometrics of the BAS3 for the eligible population in this trial, TLs administered BAS3 tests to a small sample of pupils (N=72) in Phase 1, with training in the BAS3 number skills subscale being delivered by Oxford University. Researchers from RAND Europe analysed the results from the pilot testing to understand if concerns that arose in the efficacy trial were present. The results of this analysis are presented in Table 6 and Figure 1.

The pilot test was administered to 72 children, and raw scores were available for 69 children,[14] while age standardised score were constructed for 53 pupils.[15] The BAS3 raw scores could range between 0 and 33, while in the sample the raw scores ranged between 4 and 32 based on data for 69 children. The mean raw score was 13.7, with a standard deviation of 5.8. Based on the items administered and the age of the children in the sample we constructed the age-standardised score that could range between 55 and 145. In the sample of 53 pupils for which an age-standardised score could be constructed, scores ranged between 76 and 126. The mean age-standardised score was 98.1, with a standard deviation of 13.6.

*Table 6. BAS3 descriptive statistics*

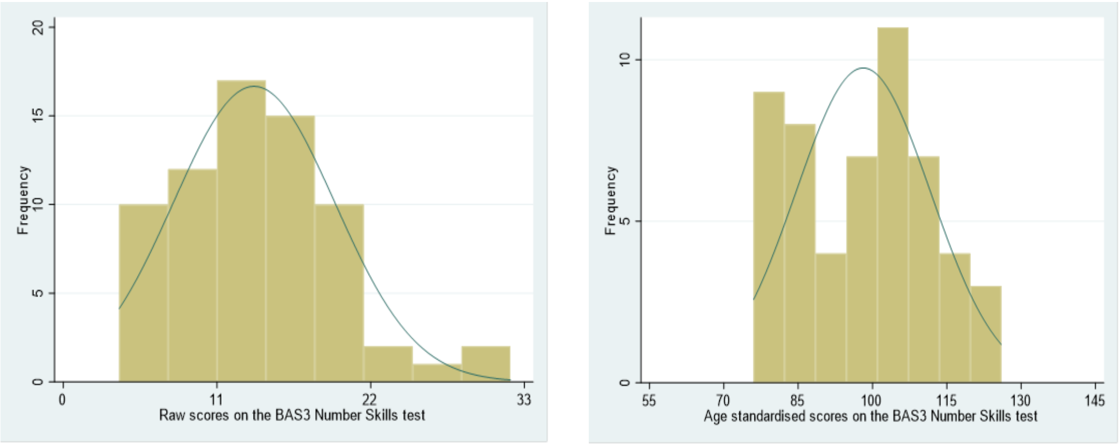
| | Raw score | Age standardised score |
|---|---|---|
| **N=** | 69 | 53 |
| **Mean** | 13.7 | 98.1 |
| **Standard deviation** | 5.8 | 13.6 |

Histograms showing the distribution of BAS3 raw and age-standardised scores are provided in Figure 1. Given the small sample size the histograms do not show perfect normal distribution, but they demonstrate that there is no evidence of ceiling or floor effects in the BAS3.

*Figure 1. Histogram showing the distribution of BAS3 raw and ability score*

Figure 2a Raw score

Figure 2b Age standardised score



Note: raw scores can take any value between 0-33. This histogram includes data for n=69/72 pupils.

Note: age standardised scores can take any value between 55-145. This histogram includes data for n=53/72 pupils.

[14] Raw scores could not be calculated for n=3 pupils as the exact item blocks that they were administered could not be determined.

[15] The lower sample size for the age standardised score was due to missing information about pupil date of birth and/or the exact item blocks that were administered being inconclusive, which meant that the age standardised score could not be calculated even if the raw score was available.