info@evidenceforlearning.org.au | evidenceforlearning.org.au | @E4Ltweets

Statistical Analysis Plan for Thinking Maths Evaluation – A teacher professional learning for middle school maths teachers (Years 6-9)

Evaluators

Australian Council for Educational Research

Prepared by Dr Hilary Hollingsworth and Dr Katherine Dix from the Australian Council for Educational Research

Intervention	Thinking Maths Evaluation
DEVELOPER	South Australian Department of Education and Child Development (DECD)
EVALUATOR	Australian Council for Educational Research (ACER)
TRIAL REGISTRATION NUMBER	ANZCTR Registry Number: ACTRN12618000437268 http://www.ANZCTR.org.au/ACTRN12618000437268.aspx
TRIAL PROJECT DIRECTOR	Dr Hilary Hollingsworth
TRIAL STATISTICIAN	Dr Katherine Dix
SAP AUTHOR	Dr Katherine Dix, Dr Hilary Hollingsworth, Mr Toby Carslake
SAP VERSION	3.0
SAP VERSION DATE	29 March 2018
E4L DATE OF APPROVAL	1 September 2018
SAP VERSIONS	1.0 - 24 November 2017

Protocol and SAP changes

This SAP refines the initial analysis plan in the Evaluation Protocol submitted 9 September 2016. The project is profiled on the Evidence for Learning website at: http://evidenceforlearning.org.au/lif/current-projects/thinkingmaths/

No substantive deviations from the original protocol have occurred.

Table of contents

Introduction

The Thinking Maths program has been developed by the South Australian Department of Education and Child Development (DECD), based on its Teaching for Effective Learning Framework. The program aims to address a significant drop in mathematics performance in NAPLAN from Year 7 to Year 9.

The program supports Year 7 and Year 8 teachers in the deep learning of mathematical content as outlined in the Australian Curriculum Mathematics. In particular, the program focuses on the following three areas for better teaching and learning of mathematics:

- Quality task design;
- Sequencing of conceptual development;
- Research-informed effective pedagogies.

Five professional learning days are run over two terms by facilitators who model rigorous teaching and learning processes, undertaking tasks with multiple entry and exit points to differentiate the curriculum to cater for students with a wide range of mathematical experience and dispositions.

After each session teachers make a commitment to implement high gain strategies to improve student achievement and engagement. Between sessions, telephone, email and online platforms support teachers' improvement efforts in a professional learning community. At each session after the first, there are three to four presentations from participants to share their experiences, successes and challenges. Figure 1 presents the program evaluation logic model for the Thinking Maths program.

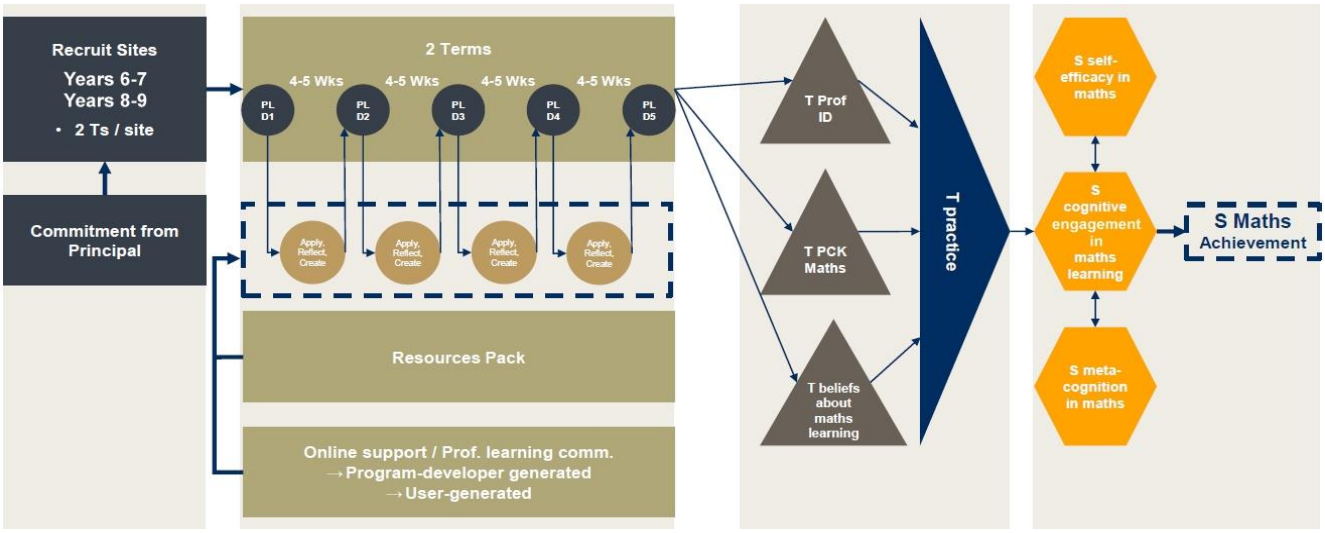

Figure 1. Program evaluation logic model for the Thinking Maths program

The program evaluator, the Australian Council for Educational Research (ACER), is the leading independent Australian organisation in educational research. The trial is structured as a clustered randomised control trial (RCT), with a focus on pairs of Year 7 and Year 8 Maths teachers recruited from 167 South Australian government schools. Although most teachers involved in the evaluation teach Years 6 to 9 (91% of students), based on teacher availability, Year levels extend from Year 5 (8%) to Year 10 (1%) and not all schools have two participating teachers. Professional learning for 120 recruited teachers from 63 schools receive the intervention and the recruited teachers from the remaining 104 schools act as a business-as-usual control. The primary outcome is improved student achievement in mathematics, and the secondary outcomes are the development of students as powerful learners of mathematics, and a shift in teachers' pedagogy towards more inclusive, student-centred learning.

The broad research question being addressed is: To what extent does the South Australian DECD Thinking Maths program improve student mathematics outcomes, and build teacher capacity to make mathematics learning deeper and more engaging?

The following five research questions underpin the impact evaluation (*if it works*).

1. Did the Thinking Maths program enable middle-school students to improve their mathematics achievement (PATMaths - Progressive Achievement Tests in Mathematics scores) above typical learning growth?
2. How did Thinking Maths develop middle-school students (Years 6-9) as powerful learners of mathematics in terms of a) mathematics self-efficacy, b) cognitive engagement in learning, and c) meta-cognition?
3. How did Thinking Maths build the capacity of teachers in terms of a) pedagogical and content knowledge, b) beliefs about mathematics teaching and learning, and c) professional identity (e.g., self-efficacy)?
4. How did Thinking Maths shift teachers' mathematics teaching practice towards a more inclusive, student-centred learning approach?
5. Did changes in teachers' practices due to Thinking Maths, influence students' mathematics outcomes?

The following seven questions underpin the process evaluation (*why and how it works*).

6. What are the critical elements of the Thinking Maths program, in terms of quality of delivery, fidelity and dosage?
7. How applicable and useful is the Thinking Maths approach (PL, online community, support, resources) in primary and secondary school settings?
8. To what extent did teachers engage with the Thinking Maths program?
9. How cost-effective is the Thinking Maths program?
10. What are the barriers and facilitators to the effective implementation of Thinking Maths in middle-school classrooms in different contexts (Year level, school socio-economic status, location, high proportions of Indigenous students)?
11. How can the Thinking Maths program be improved?

12. What are the risks and challenges in expanding the Thinking Maths program to scale?

Participant eligibility

Recruitment through an Expression of Interest was undertaken at the school level by DECD in order to attract at least 150 sites, where 63 sites received the intervention (Group A) and remaining sites acted as control (Group B). Group B included additional schools to allow for control attrition. It was preferable to have two teachers from each school receiving the intervention (120 teachers), but not to the exclusion of small schools. Eligible schools met the following criteria:

- Government school located in South Australia.
- School caters for students in Years 6-7 and/or Years 8-9 (K-12 Area schools are counted as one site).
- The teacher preferably teaches a Year 6, 7, 8 and/or 9 class in mathematics, but not to the exclusion of small or remote schools. Teachers of Year 5 and Year 10 students will not be excluded.
- The teacher has not previously participated in the Thinking Maths (or equivalent) intervention.

In order to participate in the program and the evaluation, schools agreed to:

- Participate in a briefing session, prior to random assignment and data collection.
- Be randomly assigned to have the Thinking Maths professional learning program in either a) the 2017 school year or b) late 2017 – 2018.
- Nominate two middle-school teachers (preferably) to participate in the evaluation, to participate in the program when it is offered, and to utilise skills from that program in their classrooms.
- Provide contact information (i.e., teacher email addresses) for all participating teachers to allow the evaluation team to send a survey link directly to the teachers.
- Allow all teachers in the study to participate in a brief teacher survey about their practices on two occasions, pre and post the intervention period.
- Allow the students of the participating teachers to complete a brief student survey about their views about learning mathematics on two occasions, pre and post the intervention period.
- Allow Group A teachers participating in the program to complete a feedback form at the end of each professional learning session.
- Send an opt-out consent form to the parents of all eligible students in the study (i.e., all of the students in the classes of nominated teachers) and record any opt-outs received so that the data for these students are not passed on to the evaluation team.
- Provide student class list information (name, Year) for all students in the study (i.e., students in the class of nominated teachers) to the evaluation team.

The intervention in brief

At the heart of Thinking Maths is five days of face-to-face professional learning that is programmed over two terms at 3-4 week intervals, with access to online support between sessions. The program is delivered by two DECD professional facilitators who are highly experienced middle-school mathematics teachers with extensive experience in teacher professional development and pre-service teacher training.

- *Learning:* Five PL days, 3-4 weeks apart, generally over two terms. In total, the program involves 30 hours of face-to-face professional learning, with an additional expectation of engagement in reading, journaling and presenting to the group. During the sessions 'high-gain strategies' are explicitly demonstrated with collaborative activities that involve reflecting, sharing, modelling, being the learner/practitioner, applying, and accessing professional resources.
- *Implementation:* Teachers use the four periods of 3-4 weeks in between the PL days to reflect on and apply program ideas in their mathematics classes. This implementation process follows a cycle of Action, Reflection and Creation. Teachers' journal and share the evidence of changing behaviours and outcomes in subsequent professional learning sessions.
- *Support:* Ongoing support and participation in online professional learning community.

An indicative breakdown of the time involved in the program by teachers and the exposure to students (based on DECD policy) is presented in Table 1. It should be noted that students in Group B will have similar exposure to mathematics but under business-as-usual conditions.

Table 1. Teacher and student potential exposure to the intervention

Participant	Activities	Time
Teacher	PL sessions (5 x 6hrs)	30 hrs
	Lesson preparation (per week)	2 hrs/wk
	Presentation to the group (once)	2-5 hrs
	One reading per session with reflection (5 x 2-3 hours)	10-15 hrs
	Participating in online community (voluntary)	varies
Student	Primary students learning numeracy and maths (per week)	5 hrs/wk
	Secondary students in maths class (per week)	3 hrs/wk

Study design

This efficacy study is based on a two armed (intervention and control) RCT with randomisation at the school level and involving quantitative pre and post data collection. A total of 167 schools were recruited through a self-selection process by submitting an Expression of Interest. These schools were randomly assigned to the intervention (Group A – 63 schools) and the control (Group B – 104 schools). Group A schools commenced in Term 1 2017, while Group B schools first acted as control and business-as-usual, prior to their delayed start at the commencement of Term 4 2017. Quantitative pre and post data collection included an assessment of mathematics achievement using PATMaths tests (pre-test in September 2016, post-test in September 2017), along with teacher and student pre and post online surveys. Group A teachers also completed five professional learning feedback forms. This data, collected by ACER, is augmented with existing student background data provided by DECD (e.g. student name, EDID, date of birth, gender, disability, ATSI, ESL, School Card status). The overarching approach, adapted from Torgerson and Torgerson[1], is shown in Figure 2.

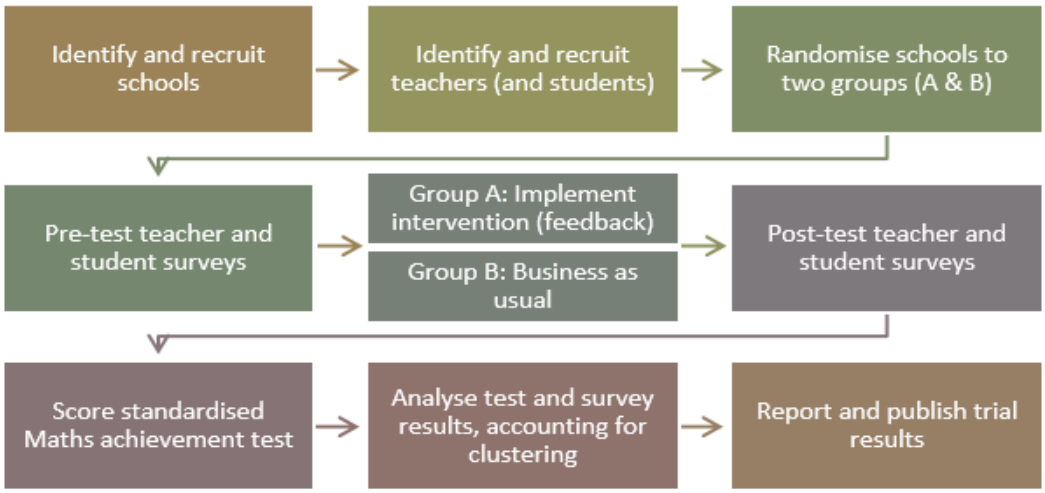

Figure 2. Key steps for school-based RCT

Sample size

In order to detect an effect that is sufficiently large to be of educational significance at the student level (i.e. above 0.2), and given that teachers are clustered within schools, the following recommendations about sample size were provided at the initial design stage. We take the desired alpha to be 0.05 and power to be 0.8, with a detectable effect size of small (Cohen's d = 0.2). Even using the simplest formula for a RCT block design comparing two groups of equal size, we also need to take into account the design effect of clustering and estimated intra-cluster correlations (ICC). Since students in one school are more like

[1] Torgerson, C.J. & D.J. Torgerson (2013). *Randomised trials in education: An introductory handbook*. London, Education Endowment Foundation.

each other than students in another school[2], the sample is not a simple random sample, and results in a net loss of information. In other words, from a statistical perspective, similarities between students in the same class effectively reduce the number of participants in the intervention[3]. The 'design effect' is used to estimate the extent to which the sample size should be inflated to accommodate for the homogeneity in the clustered data. In line with similar studies in Australia, Zopluoglu's[4] recommendation of an Australian ICC coefficient range of 0.2-0.3 (p.264), and the PISA 2012 Technical Report[5] with an Australian ICC for mathematics of 0.28 (p.439), we initially estimated an ICC coefficient of $\rho = 0.3$, but will review this once data are available.

In order to minimise sample size and achieve the desired Minimum Detectable Effect Size (MDES) of 0.2, the Bloom MDES formula with both level-1 and level-2 covariates[6] was used, which increases the power of a cluster-level RCT by including pre/post-test correlation. The hierarchical model controls for the majority of variance, which is known to be explained by prior achievement, both at the school level and the student level. The remaining variance, therefore, is more sensitive to explaining the impact by teacher participation (or not) in the intervention.

Accordingly, a minimum sample of 120 schools (60 intervention, 60 control) was needed to achieve a MDES of 0.2 with covariates that accommodate design effects and provide allowances for participant attrition and missing data. Through the recruitment process, a sample of 167 schools was achieved. Participant attrition, likely to be higher in the control group, will counter possible effects of unequal cluster size[7].

Randomisation

This study used concealed randomisation so that there was no foreknowledge of the randomised allocation[1]. Randomisation at the school level was done after teachers had been recruited and consented to participate in the study. This occurred after the briefing session and once they had completed the teacher pre-survey. Accordingly, all participants, including teachers, schools and DECD (the recruiters and program implementers) did not know which group the schools were randomised into until their baseline pre-survey had been submitted. In order to maintain independence and concealment from the program implementers (DECD) and the evaluation funders (SVA), DECD provided the sampling frame of participating schools to ACER, and ACER undertook a simple random sample of cases on the de-identified list of schools in which only a school ID number and the number of participating

[2] Hutchison, D. & Styles, B. (2010). A Guide to Running Randomised Controlled Trials for Educational Researchers. Slough: NFER.

[3] Torgerson, C.J. & Torgerson, D.J. (2013). *Randomised trials in education: An introductory handbook*. London, Education Endowment Foundation.

[4] Zopluoğlu, C. (2012). A cross-national comparison of intra-class correlation coefficient in educational achievement outcomes. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *3*(1), 242-278.

[5] OECD (2014). PISA 2012 technical report. OECD, Paris.

[6] Bloom, H.S., Richburg-Hayes, L. & Rebeck Black, A. (2007). Using covariates to improve precision for studies that randomise schools to evaluate educational intervention. *Education Evaluation and Policy Analysis*, 29 (1), 30–59.

[7] Eldridge, S. M., Ashby, D., & Kerry, S. (2006). Sample size for cluster randomized trials: effect of coefficient of variation of cluster size and analysis method. *International Journal of Epidemiology*, *35*(5), 1292-1300.

teachers was known. Specifically, the 'Select cases: random sample' dialog box in SPSS 22 was used to select an exact number of cases from the total list. The trial statistician performed the random sample once a colleague, independent to the project, had de-identified the list and assigned the school ID number. Accordingly, the trial statistician was blind to the schools being randomly allocated to either the treatment or control groups during the randomisation procedure.

Because the developers (DECD) budgeted for an allocation of 120 teachers to undertake the Thinking Maths Professional Learning sessions in 2017, it was necessary to achieve a random sample of schools that resulted in 120 teacher places. However, while most of the 167 participating schools nominated two teachers, as requested, a small number of schools only nominated one teacher, and two K-12 schools nominated four teachers (two in the primary year-levels and two in secondary). A total of 318 teachers were nominated by schools. Several simple random samples were drawn, first by specifying 60 schools (which yielded fewer than 120 teacher places), and then repeating the process up to 63 schools, at which point the desired number of 120 teacher places was achieved. This group of 63 randomly sampled schools and their 120 teachers formed the intervention group (Group A). The remaining 104 schools and their teachers formed the control group (Group B).

It should be noted that while no nominated teachers had previously undertaken Thinking Maths, 26 participating schools had other teachers complete Thinking Maths in 2014-2016. Nine of these schools were in the intervention group and 17 schools were in the control group.

Participation and follow-up

While data collection and cleaning is still in progress, a preliminary CONSORT flow-diagram[8] is displayed in Figure 3 to provide an indication of participation and the extent of missing data.

Outcome measures

Primary outcome

The primary outcome identified in this evaluation – the outcome that determines whether or not the intervention is effective – is **improved student achievement in mathematics for all learners.** This is measured by the *ACER Progressive Achievement Tests in Mathematics* (PATMaths) test, routinely completed by all students in South Australian government schools since 2015.

PATMaths[9] is a thoroughly researched, Australian test designed to provide objective, norm-referenced information to teachers about the level of achievement attained by their students in the skills and understanding of mathematics. Each of the ten PATMaths tests assesses the

[8] Moher, D., Hopewell, S., Schulz, K. F., Montori, V., Gøtzsche, P. C., Devereaux, P. J. & Altman, D. G. (2010). CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. *Bmj*, *340*, c869.

[9] For further information about ACER PATMaths visit https://www.acer.edu.au/patmaths

content of one year level of the Australian mathematics curriculum from Year 1 to Year 10, which assumes coverage of the curriculum of lower Year levels. All PATMaths tests have a common achievement Rasch scale, enabling results to be compared between different Year levels. The PATMaths Fourth Edition tests (2013) cover six mathematics strands, namely, Number, Algebra, Geometry, Measurement, Statistics, and Probability. Each test comprises at least five items for each of the strands it covers with a total of 40-50 items depending on the Year level. Within a test, the items are ordered from easiest to most difficult. The test is completed by students online within 40 minutes.

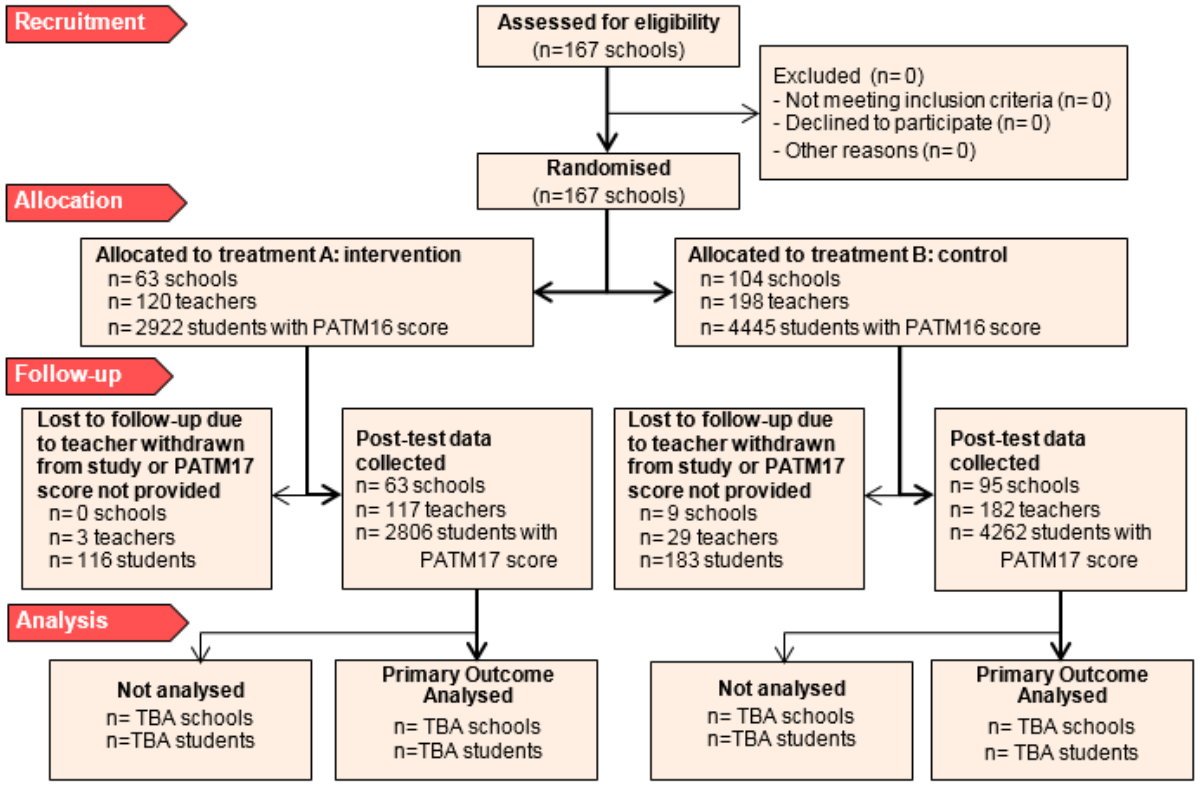

Figure 3. Participation flow diagram

In accordance with DECD's mandatory annual administration of PATMaths, the pre-test data was collected during September 2016, with data retrieved retrospectively once the schools, teachers and their students participating in the Thinking Maths evaluation were identified. The post-test data collection occurred in September 2017.

The test is scored instantaneously through the ACER Test Scoring and Analysis software. The PAT *raw score* is the number of correct answers on a test. The PAT *scale score* is the test raw score converted to the relevant PAT scale. Based on analysis of the data using the Rasch model, this scale enables student achievement and question difficulties to be located on the same scale across Year levels. The standardised PAT scale score for middle-school cohorts generally range between 50 to 200 scale units and is the primary outcome measure used in this evaluation. A positive pre-post coefficient difference indicates learning growth.

DECD is providing ACER with the pre (2016) and post (2017) PATMaths scale scores for each participating student. The resulting database will be coded and de-identified following data-linkage of the PATM16 (prior achievement) and PATM17 (primary outcome) scale data and the student pre-survey (prior attitudes) and post-survey (secondary outcomes) data. Reflecting the nested nature of the data, students will also be linked to classes (teachers) and schools.

Secondary outcomes

The secondary outcomes identified in this evaluation broadly align to the original Evaluation logic model for the Thinking Maths program (see Figure 1). Given the lack of program documentation, observations of the Thinking Maths professional learning days by ACER staff the previous year strengthened an understanding of what needed to be assessed. Scales and items were designed with pre-post capacity in mind and to be appropriate for participants in primary or secondary settings, as well as those either in the control or treatment groups. As part of the process of survey design, a review of the literature regarding attitudes and beliefs towards mathematics teaching and learning was conducted to source candidate scales and items for the instruments (e.g. PEEC, 2016; PISA 2012; Kong, 2003; Fredericks et al., 2005; Pintrich & DeGroot, 1990; Deci et al., 2013; LSAC, 2013; DECD, 2016; KidsMatter, 2014). Where necessary, items were modified or new items were developed to meet the specific needs of the evaluation. Confirmatory factor analysis and Item reliability analysis were conducted in order to summarise the items and derive meaningful constructs in the form of mean scores. Reliabilities of 0.80 or more are described as high; between 0.70 and 0.80 as moderate; and between 0.60 and 0.70 as low.

The four secondary outcomes collected through the Student Survey will be:

- **Students' mathematics anxiety and low self-concept** (SASE): Mean score of 10 items measured on a five-point Likert scale of Strongly disagree (1) to Strongly agree (5), with high internal reliability (α=0.89).
    - I get nervous doing maths problems
    - I get worried when I have to do maths homework
    - Maths is hard to understand so I just try to learn the steps
    - I worry that I will get poor grades in maths
    - I don't like it when the teacher asks me questions in maths
    - I usually do well in maths (reverse scored)
    - I am just not good at maths
    - I learn things quickly in maths (reverse scored)
    - I am good at working out difficult maths problems (reverse scored)
    - Maths is harder for me than any other subject

- **Students' cognitive engagement** (SCOG): Mean score of 5 items measured on a five-point Likert scale of Strongly disagree (1) to Strongly agree (5), with high internal reliability (α=0.81).
    - I know what my teacher expects me to do
    - I am interested in what my teacher says
    - My teacher is easy to understand

  - My teacher believes all students can be good at maths
  - My teacher gives me interesting things to do in maths

- **Students' metacognitive strategies** (SMET): Mean score of 5 items measured on a five-point Likert scale of Strongly disagree (1) to Strongly agree (5), with moderate internal reliability (α=0.76).
  - I check my maths school work for mistakes
  - I try to connect the things I am learning in maths with what I already know
  - I like to try and use what I learn in maths, in real life
  - When I become confused about something in maths, I go back and try to figure it out
  - I make up my own maths problems to test my understanding

- **Students' learning through effective teaching practice** (SETL): Mean score of 16 items measured on a five-point Likert scale of Never (1) to Always (5), with high internal reliability (α=0.89).
  - My teacher asks me or my classmates to present our mathematical thinking
  - My teacher asks questions to check whether we have understood what was taught
  - At the beginning of a maths lesson, the teacher reminds us of what we did in the previous lesson
  - My teacher asks me to explain my answers
  - Our maths assignments and homework make me think hard about what I'm learning
  - My teacher makes sure I understand before we move on to the next topic
  - My teacher gives different work to classmates depending on their ability (e.g. to those who have difficulties learning or to those who finish quickly)
  - We work in groups to come up with joint solutions to a problem
  - My teacher asks us to guess or estimate the answer to a problem before we calculate it
  - My teacher asks me questions in maths that challenge my thinking
  - My teacher tells me about how well I am doing in my maths class
  - My teacher talks to me about what I need to do to become better in maths
  - My teacher shows an interest in every student's learning
  - My teacher gives extra help when students need it
  - My teacher continues teaching until we understand
  - My teacher gives students an opportunity to express opinions and ask questions

The three secondary outcomes collected through the Teacher Survey will be:

- **Teacher professional identity and self-efficacy** (TPID): Mean score of 7 items measured on a five-point Likert scale of Not at all (1) to A great deal (5), with high internal reliability (α=0.89).

When teaching maths, to what extent can you do the following:

- Engage all students
- Help your students think critically
- Improve the understanding of a student who is failing
- Motivate students who show low interest in maths
- Help your students value maths learning
- Help students to believe they can do well in maths
- Create opportunities for all students to experience productive struggle

- **Teacher pedagogical and content knowledge** (TPCK): Mean score of 10 items measured on a five-point Likert scale of Not at all (1) to A great deal (5), with high internal reliability (α=0.91).

  How confident are you in the following areas:

  - Designing learning with the Australian Curriculum mathematics proficiencies
  - Designing learning with the Australian Curriculum mathematics content
  - Knowing the mathematics developmental learning progression across Years 6 and 9
  - Differentiating your teaching of the Australian Curriculum Mathematics
  - Creating and maintaining a mathematical learning environment that challenges all students
  - Creating and maintaining a mathematical learning environment that supports creative and critical thinking
  - Using questioning to develop students' conceptual understanding
  - Using questioning to diagnose students' conceptual misunderstandings
  - Identifying students' learning challenges
  - Providing timely feedback to students

- **Teacher beliefs about mathematics learning** (TBEL): Mean score of 3 items measured on a five-point Likert scale of Strongly disagree (1) to Strongly agree (5), with low internal reliability (α=0.68).

  - I deeply believe that everyone can learn maths
  - You are either good at maths or you're not (reverse scored)
  - Some students are probably never going to be good at maths (reverse scored)

The Teacher and Student Surveys were administered online and were designed to take no more than 30 minutes to complete (within a lesson time). Participating teachers were provided with links to the surveys through an email invitation with instructions to complete the Teacher survey and administer the Student survey. Survey administration occurred on two occasions. The pre-surveys were conducted in February 2017 before recruited schools were given their random allocation to the control or treatment groups. The post-surveys were conducted in October 2017 following the PATM administration period. Responses were automatically scored and collated into a secure downloadable database through the online survey hosting platform. Pre and post survey data were cleaned and, along with the PATM

data, matched using student class lists, preserving the nestedness of students and teachers in classes in schools, at which point the data was de-identified.

Analysis

This section discusses in more detail the plan for how the analysis of the data collected is defined and treated in order to address the aims of the evaluation. Analytical methods have been selected to reflect the study design, randomisation choices, and the nested structure of educational data. The section begins with general discussion about our approach to analysis and is followed by the measurement framework that summarises the specific aspects being measured and their treatment.

Primary intention-to-treat (ITT) analysis

In accordance with the guidance, the analysis of primary and secondary outcomes measures will be undertaken on intention to treat basis meaning that all those allocated to treatment and control in the randomisation are included.

The primary student outcome measure, student mathematics achievement (PATM17), will be analysed using a hierarchical linear model (HLM) to reflect the nested nature of the data and the method of assignment, with students nested within classes, within schools. The student model will include individual student's prior PATM16 score. Intervention and control groups will be compared by including an intervention indicator at the school level, where Intervention = 1 and Control = 0. Figure 4 presents a standard two-level hierarchical model of the factors influencing students' mathematics achievement.

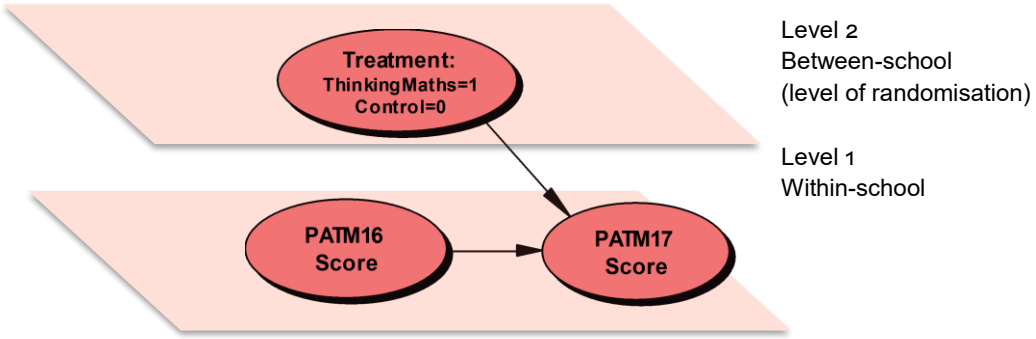

Figure 4. The standard two-level HLM model of factors influencing mathematics achievement

Analysis of whether the intervention was effective or not will be based upon the following standard HLM regression model:

$$PATM17 = \beta_{00} + \beta_{01}*Treat + \beta_{10}*PATM16 + \varepsilon$$

where:

PATM17 = students' PATMaths 2017 post achievement score (independent of Year level).

PATM16 = students' PATMaths 2016 prior achievement score (independent of Year level).

Treat = a binary variable indicating whether the student was enrolled in an intervention (1) or and control (0) school.
ε = error term (with student clustered within school)

The direct effect of Treat on PATM17 will test if there is a significant difference in the maths achievement outcomes of students in the intervention schools compared to students in the control schools, controlling for prior achievement.

Interim analyses

Due to the short time-span of this trial, no interim analyses is required.

Imbalance at baseline for analysed groups

At the time of school assignment, only school-level information and the number of nominated teachers in each school was available. We will check balance in baseline student characteristics; specifically, student PATM16 scale scores that are used as the pre-test control in this trial. Imbalance in PATM16 scores will be tested in the HLM model. The direct effect of Treat on PATM16 (i.e. intercept) will test if there is a significant difference in the maths achievement at baseline in the two groups of schools. Differences in test scores will be reported as effect sizes. While there may be imbalance due to the random nature of assignment, we do not expect imbalance at baseline due to the low attrition of schools (4% of recruited schools were lost to attrition). However, if the data are found to be substantially imbalanced between the intervention and control groups at baseline, additional sensitivity analysis will be performed to understand the nature of the imbalance. If needed, this analysis will consider the impact of missing data (completers only vs imputed missing data)[10]. Balance comparison of observables, in terms of school and student characteristics, will be reported (see Table 6).

Missing data

As with any data collection process, missing data may arise for several reasons:

- The participant might chose not to answer an item or inadvertently miss an item.
- The participant might feel that a section of items is not relevant to them personally.
- The survey or feedback form is returned only partially completed, not completed at all, or not administered.

Missing data will be coded with a single missing code value (-999) to represent all missing data. When data are missing, the power of the analysis to detect statistically significant effects is reduced and, depending on the mechanism by which data are missing, the estimated effects and standard errors can potentially be biased. Methods will be used during scale score construction that avoid the need to impute missing data. Careful consideration will be given to the existence of missing data likely to bias the findings of the evaluation with respect to its representativeness.

[10] de Souza, R. J., Eisen, R. B., Perera, S., Bantoto, B., Bawor, M., ... & Thabane, L. (2015). Best (but oft-forgotten) practices: sensitivity analyses in randomized controlled trials. *The American journal of clinical nutrition*, *103*(1), 5-17.

If 5% or less of students have incomplete post-test outcome data, analysis omitting these students will be conducted. In other words, analysis using listwise deletion of any student with incomplete information will be conducted. Previous research has found that when 5% of the data are missing, bias is low across the various approaches to handling missing data in analysis including listwise deletion[11].

If greater than 5% of the students have incomplete outcome data, the mechanism by which data are missing will be investigated using Little's[12] MCAR test and multiple imputation accounting for the nested structure of the data will be considered as an alternative to listwise deletion if data are not missing at random. The amount and pattern of missingness will be reported.
With regard to the primary outcome measure provided retrospectively by DECD, the possible reasons why students might be missing data in this study include:

- DECD was unable to match the participating student in the PATMath database to the class list generated 6 months earlier (match on student name, teacher, school).
- Student was absent on the day of testing.
- Student was not tested due to having special educational needs or disability.

Non-compliance with intervention

Given the complexity of implementing a program in schools, it is anticipated that some teachers will engage more readily than others with the Thinking Maths program, and so will be better able to effect change. We will conduct a Complier Average Causal Effect (CACE) analysis. Compliance will be a numerical score, based on the number of training sessions each teacher has attended (scored 0 to 5). The responses to the PD feedback form, collected at the end of each training session during the evaluation period, will be used to verify the participation data (sign-on sheets).

Secondary outcome analyses

The four student indices and three teacher indices derived from the pre-post surveys, as listed below, will be analysed using the same model as described above in the primary ITT analysis (see Table 8).

- Students' mathematics anxiety and low self-concept (SASE)
- Students' cognitive engagement (SCOG)
- Students' metacognitive strategies (SMET)
- Students' learning through effective teaching practice (SETL)
- Teacher professional identity and self-efficacy (TPID)

[11] Puma, M.J., Olsen, R.B., Bell, H.S., & Price, C. (2009). *What to Do When Data Are Missing in Group Randomized Controlled Trials* (NCEE 2009-0049). Washington, DC: National Center for Education Evaluation and Regional Assistance, Institute of Education Sciences, U.S. Department of Education.
[12] Little, R. J. (1988). A test of missing completely at random for multivariate data with missing values. *Journal of the American Statistical Association*, 83, 1198-1202.

- Teacher pedagogical and content knowledge (TPCK)
- Teacher beliefs about mathematics learning (TBEL)

Additional analyses

No additional outcomes analyses are expected for this study. During the exploratory analysis phase, we will estimate our primary and secondary analyses HLM models including additional student level covariates (e.g. gender, ATSI).

Subgroup analyses

School Card: Subgroup analysis will be conducted for the population of School Card students. For this analysis, the primary and secondary outcome analysis models will be re-estimated using data limited to this sample (Student School Card status: 0=non-recipient; 1=recipient). Furthermore, an interaction model will be run on the entire data. This will mirror the main primary outcome model but also include School Card and the interaction between School Card and group as covariates.

School type: Similarly, subgroup analysis will be conducted for the population in Primary (Years 5-7) and in Secondary (Years 8-10) schooling contexts using data limited to each sub-sample.

Effect size calculation

Primary outcome results will be reported as scale scores as well as effect sizes that standardise the estimated impacts. The numerator will be the regression adjusted estimate of the impact of Thinking Maths from the multi-level model and the denominator will be the standard deviation of the outcome for the full sample.

When determining the effect size we will use the total variance, rather than the residual variance from the clustered model. Variations in a post-test outcome, due to different sources, must be fully accounted for in a statistical model [13,14,15]. For cluster randomised trials, the total variability can be decomposed into random variation between students (σ_i) and heterogeneity between schools (σ_s). Effect sizes for cluster randomised trials with equal cluster size and using total variance are calculated as:

$$Effect\ Size = \frac{(\overline{Y_T} - \overline{Y_C})}{\sqrt{{\sigma_i}^2 + {\sigma_s}^2}} = \frac{\beta_{Treat}}{\sqrt{{\sigma_i}^2 + {\sigma_s}^2}}$$

[13] Per EEF guidance, see https://educationendowmentfoundation.org.uk/public/files/Evaluation/Analysis_for_EEF_evaluations_REVISED_Dec_2015.pdf.

[14] Tymms, P. (2004) Effect sizes in multilevel models. In I. Schagen & K. Elliot (Eds.) *But what does it mean? The use of effect sizes in educational research,* pp.55-66. Slough: National Foundation for Educational Research.

[15] Hedges, L. V. (2007). Effect sizes in cluster-randomized designs. *Journal of Educational and Behavioral Statistics*, 32(4), 341-370.

We will calculate the effect of the program as shown in the example in Table 7 (see section: Report tables). An equivalent table for the secondary outcomes will also be produced.

Report tables

The guidelines require several pre-specified tables. Table 2 shows the provision of five key conclusions from the evaluation.

Table 3 shows an example table of a summary of the impact on primary outcomes based on E4L cost evaluation guidance. A description of the other tables will be provided as confirmed.

Table 2. Key conclusions from the Thinking Maths trial

Key conclusions
1. Headline for schools …
2.
3. Important factors for implementation …
4.
5. Possible further research

Table 3. Summary of impact on primary outcome

Group Intervention (A) vs Control (B)	Effect size (95% CI)	Estimated months' progress	E4L security rating	E4L cost rating
All A vs B PATM17				
School Card: A vs B PATM17				
Non-School Card: A vs B PATM17				
Primary: A vs B				
Secondary: A vs B				

Table 4. Evaluation timeline

Date	Activity

Table 5. Minimum detectable effect size at different stages

Stage	N [schools] (n=intervention; n=control)	Proportion of variance in outcome explained by pre-test + other covariates	ICC	Power	Alpha	Minimum detectable effect size (MDES)
Protocol	167 (63; 104)	TBA	0.3	0.80	0.05	0.2
Randomisation	167 (63; 104)	TBA	0.3	0.80	0.05	0.2
Analysis						

Table 6. Balance comparison of observables – school and student characteristics

Variable	Intervention - Group A		Control - Group B	
School-level	**n (missing)**	**Mean or %**	**n (missing)**	**Mean or %**
School size				
School type (% primary)				
School location (% metro)				
Average SES (category)				
% ATSI in the school				
Student-level (categorical)	**n (missing)**	**Percentage**	**n (missing)**	**Percentage**
School Card holder				
Gender (% Male)				
Year level				
Student-level (continuous)	**n (missing)**	**Mean**	**n (missing)**	**Mean**
Scale score PATM16				
Student age				

Table 7. Primary analysis

Outcome	Raw means				Effect size		
	Intervention (Group A)		Control (Group B)				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (A;B)	Hedges g (95% CI)	p-value
PATMath17							

Table 8. Secondary analysis

Outcome	Raw means				Effect size		
	Intervention (Grp A)		Control (Grp B)				
	n (missing)	Mean (95% CI)	n (missing)	Mean (95% CI)	n in model (A;B)	Hedges g (95% CI)	p-value
Students' mathematics anxiety and low self-concept (SASE)							
Students' cognitive engagement (SCOG)							
Students' metacognitive strategies (SMET)							
Students' learning through effective teaching practice (SETL)							
Teacher professional identity and self-efficacy (TPID)							

Teacher pedagogical and content knowledge (TPCK)							
Teacher beliefs about mathematics learning (TBEL)							