

Statistical Analysis Plan

QuickSmart Numeracy Evaluation – A small-group intervention to increase students' maths fluency and automaticity

Evaluators

Teachers and Teaching Research Centre (TTRC), University of Newcastle

Prof Jenny Gore, Dr Drew Miller, Dr Elena Prieto-Rodriguez, Dr Jess Harris, Dr Adam Lloyd, Dr Leanne Fray

Intervention	QuickSmart Numeracy Evaluation
DEVELOPER	SiMERR University of New England
EVALUATOR	University of Newcastle
TRIAL REGISTRATION NUMBER	ACTRN12616001612404
TRIAL STATISTICIAN	Dr Andrew Miller
TRIAL CHIEF INVESTIGATOR	Laureate Prof Jenny Gore
SAP AUTHOR	Andrew (Drew) Miller, Jenny Gore, Jess Harris, Elena Prieto-Rodriguez, Leanne Fray, Adam Lloyd, Wendy Taggart
SAP VERSION	2.0
SAP VERSION DATE	22 January 2019
E4L DATE OF APPROVAL	1 April 2019

Protocol and SAP changes

1. Per-protocol analysis was not included in the original protocol document, and is included in SAP version 1 to assess treatment effects in the presence of non-compliance.
2. Primary outcome administration – Progressive achievement tests were originally to be administered electronically (online version). Logistics issues and the relatively small cohort at each school meant paper-based administration was more efficient in this situation.
3. Secondary outcome removal - The secondary outcome for mathematics achievement (NAPLAN numeracy) reported in SAP version 1 has been removed from the protocol. Due to the time frame between repeated measures data (2 years), this data is considered susceptible to bias introduced by the student exposure to multiple teachers across this time period.
4. Primary analysis approach changed from a Linear Mixed Model to Repeated Measures (ANCOVA adjusted for pre-test score and accounting for clustering within class) procedure in SAP version 2.
5. Instrumental Variable analysis replaces per-protocol and regression of intervention group only in SAP version 2 for the treatment effects in the presence of non-compliance.

Note. Per-protocol analysis at 75% and 90% of compliance were outlined in SAP version 1, and analysis reported in the original version of the report. These analyses were removed to comply with the suggestions of the initial review, and were replaced with an instrumental variable analysis for the purposes of technical correctness.

Unless stated within this analysis plan, data handling methods are considered A priority.

Table of contents

Protocol and SAP changes	2
Table of contents	3
Introduction	4
Intervention	4
Study design	4
Randomisation	5
Sample size	6
1. Calculation for repeat measures design (ANCOVA)	6
2. Adjustment for clustering	6
3. Attrition	7
Follow-up	9
Outcome measures	10
Primary outcome	10
Secondary outcomes	10
Analysis	11
Primary analysis	11
Interim analyses	11
Imbalance at baseline for analysed groups	12
Missing data	12
Secondary outcome analyses	13
Treatment effects in the presence of non-compliance	13
Subgroup analyses	14
Effect size calculation	14

Introduction

QuickSmart Numeracy ('QuickSmart') is an intensive 30-week tutoring intervention developed by the SiMERR National Research Centre. The intervention aims to increase fluency and automaticity in mathematics for Year 2 to 10 students (aged between approximately 6 to 16 years) performing in the bottom third of their national cohort in mathematics. Pairs of students are withdrawn from class for three 30-minute sessions a week which are delivered by a trained school staff member (instructor), typically a teacher's assistant. This document outlines the planned analysis of a two-arm randomised controlled trial comparing intervention and control condition students from within the same class group. Primary (Year 4) and Secondary (Year 8) students from 24 schools (12 per cohort) based in Sydney New South Wales were recruited for this evaluation. This plan outlines the study design, randomisation procedure, sample size calculation, and an intervention recruitment and retention report. It also outlines our primary and secondary outcomes analyses, effect size calculation, missing data procedures and sub-group analyses.

Intervention

QuickSmart Numeracy is an intensive 30-week tutoring intervention which aims to increase fluency and automaticity in mathematics. Pairs of students are withdrawn from class for three 30-minute lessons a week, usually in a dedicated room within the school, with access to computers. Sessions are typically overseen by teacher assistants, and sometimes teachers or school executive members. The QuickSmart instructional approach focuses on the role of automaticity in developing students' fluency and facility with basic academic facts, and is informed by relevant literature associated with learning difficulties/disabilities and quality instruction (Baker, Gersten, & Lee, 2002; McMaster, Fuchs, Fuchs, & Compton, 2005; Westwood, 2007), effective instruction (Rowe, Stephanou, & Urbach, 2006), mathematics education (Fuchs & Fuchs, 2001) and educational interventions (Desh-ler, Mellard, Tollefson, & Byrd, 2005; Marston, 2005).

Study design

QuickSmart was evaluated using a multi-site two-arm individual randomised trial comparing intervention and control condition students from within the same class group. Intervention students within a class received the QuickSmart program in addition to regular classroom tuition, while the wait-list control condition students in the same class received only their regular classroom tuition. This design provided an intervention and control condition within each class group, accounting for the effect of the classroom teacher on students in both conditions, and enabling evaluation of the effect of the QuickSmart intervention in addition to regular classroom mathematics instruction.

Primary (Year 4) and Secondary (Year 8) school cohorts were recruited for the evaluation. Students identified in the bottom 30% of the most recent NAPLAN round (Year 3 for Year 4

cohort, Year 7 for Year 8 cohort), with no existing diagnosis of a learning disorder were eligible to participate. The recruitment target was 24 schools (12 Primary / 12 Secondary). This design was based on an approximation of 12 students from each Primary school, and 16 students from each Secondary school (regardless of the number of classes). In total, 23 schools were recruited, resulting in 146 intervention students and 142 wait-list control students (n = 288) from 70 class groups after randomisation. Details of each cohort are recorded below:

Primary: 12 schools; 30 classes; 67 intervention / 66 control (n = 133)

Secondary: 11 schools; 40 classes; 79 intervention / 76 control (n = 155)

All participating students completed baseline measures prior to randomisation during Term 1 (March), 2017. Students in both conditions received regular classroom mathematics tuition throughout 2017, and students allocated to the intervention condition participated in the QuickSmart program for the 2017 academic year. Follow-up measures were collected at the end of Term 4 (November), 2017, and again in Term 2 (May), 2018 at 6-months post intervention.

As QuickSmart students received more frequent numeracy instruction than those in the control group in 2017, the 6-month follow-up assessment (reported here) was used to assess the effects beyond immediate exposure to the QuickSmart intervention. Results from the immediate follow-up time point (November 2017) are reported as an interim analysis in the results below for the primary outcome to give an indication of the immediate intervention effects. Students allocated to the wait-list control group were offered the opportunity to participate in the QuickSmart program from Term 2, 2018, after all intervention data was collected.

Note. Reporting of the immediate follow-up time point (November 2017) was not specified in the original protocol.

Randomisation

Randomisation occurred after baseline assessment and took place in March 2017, using the non-scaled version of the primary outcome variable (PAT-M raw score). Randomisation was undertaken at the individual level within each class group, resulting in an intervention and a control group within each class group. Participants within each class were stratified by gender, and ranked within strata by their baseline PAT-M raw score. Gender based pairs were formed using their respective ranking (e.g., 1 and 2, & 3 and 4, etc). The highest ranked participant within a pair was randomised to one of two conditions first (intervention or 18-month wait-list control), with the other participant in the pair allocated to the alternate condition. A member of the University of Newcastle evaluation team (AM) carried out the randomisation procedure by tossing a coin. During randomisation, “heads” resulted in allocation to the intervention group, and “tails” the control group. The evaluator carrying out

the randomisation procedure recorded the outcome against the name of each person within a class and the trial manager was responsible for communicating the results with the relevant school contact.

In the case of an uneven number of students within gender strata from an individual class (e.g., 3 girls and 5 boys), any participants not in a pair were randomised individually to one of the two conditions via a coin toss.

Sample size

The calculated sample size included considerations of statistical power and access to a convenient sample indicated by the intervention research team. Sampling was calculated on a per-cohort basis (Primary and Secondary) to ensure sufficient recruitment for analysis to take place for each cohort. The sample size calculation was based on a three-step process:

1. Calculation for repeat measures design (ANCOVA)

G*Power (version 3.1.9.2) was used to determine an unadjusted sample for the desired effect, type 1 error and power. An ANOVA F test (Repeated measures, within-between interaction) was used as the correlation among repeated measures could be taken into consideration during calculation. Assumptions: Effect $f = 0.15$, Alpha = 0.05, Correlation among measures = 0.5, Power = .80

The estimated effect of 0.15 is considered small-to-medium within an F-test (Cohen, 1992). As the lowest 30% of NAPLAN achievement were recruited, a low correlation among repeated measures was used as a low pre-post correlation was expected due to the lack of variance available in this group. A sample of 90 students was required for the unadjusted (for clustering) cohort.

2. Adjustment for clustering

Clustering in educational studies has the effect of reducing the amount of data available during analysis (Dreyhaupt, Mayer, Keis, Öchsner, & Muche, 2017). The more students with groups (e.g., within schools) are like each other, and the more their groups differ from each other (e.g., between schools), the closer each group moves to acting like a single data point during analysis, reducing the power available. To adjust for clustering, the correction factor $[1 + (m - 1) \times ICC]$ was applied (Donner & Klar, 2000), where m = students per school and ICC = the intra-class correlation coefficient (between school variance / (between school variance + within school variance). Assumptions: $\rho(w) = 0.05$ (correlation coefficient for within cluster variation), subjects per cluster (school) = 10.

The PISA 2012 Technical Report suggests an Australian ICC for mathematics of 0.28 (OECD 2014, p.439). However, based on data from Lamb & Fullarton (2001) demonstrating between school variance of 0.104 when between class variance of 0.279 is taken into account at the second level of analysis (resulting school level ICC = 0.1), an ICC of 0.05 was chosen for two

reasons: 1) the cohort was being randomised at the individual level within clusters (initially believed to be schools), so clustering was expected to have less of an effect within the analysis; and 2) restricting the cohort to the lowest achieving third of students nationally we believed would likely produce less variation between clusters, thus reducing the ICC. A correction factor of 1.45 resulted in an adjusted student sample of 131 students per cohort (Primary and Secondary).

3. Attrition

To account for the loss of students across the trial period (e.g., moving schools, dropout etc), an arbitrary value of 5% was added to the adjusted sample. The resulting sample was 137 students to be recruited from each cohort (Primary and Secondary).

Through the recruitment process, 135 students (67 intervention / 68 control) from 30 classes at 12 schools were recruited for the Primary cohort, and 169 students (85 intervention / 84 control) from 40 classes at 11 schools for the Secondary cohort. Table 1 presents the Minimum Detectable Effect Size (MDES) at different stages of the study for the combined cohort and Primary and Secondary subgroups. As students were randomised within multiple class groupings at each school (rather than within one class at the school), the class a student belonged to was used to examine the effect of clustering within the sample. Linear mixed models used for outcome analysis also used the class a student belonged to for adjustment of clustering within models (see below).

A variance components model with the student (level 1) and the class (level 2) was used to establish the proportion of variance within the outcome variable (PAT-M) attributed to clustering (variance not explained at the student level that contributes to the ICC). This was considered a conservative approach as all variance not attributed to the student (class and school) was likely contained in the ICC when using the student's class, resulting in a higher ICC (and larger correction factor) than using the school as the cluster level. To maintain the most transparent research process, MDES was not modified using covariates (Bloom, Richburg-Hayes, & Black, 2007).

Minimum detectable effect size was determined by back transformation of the recruited sample given the available information (pre-post correlation, ICC, cluster number and average cluster size) using the process described above. The back transformed raw sample (total sample / correction factor calculated using available information) was used in a sensitivity analysis in G*Power to determine the resulting effect size f . This effect size f was converted to effect size d for easier interpretation using the calculations in Cohen (1988).

Clustering at the class level produced higher ICC values than expected, however this was somewhat counteracted by a reduction in the average cluster size when calculating the correction factor. The MDES for the Primary and Secondary subgroups were considered slightly underpowered ($d > 0.2$), however when the cohorts were combined for the total analysis, the sample was considered sufficient to detect an effect as low as $d = 0.2$ when the ICC was produced using variance components of the final assessment value (PAT-M time 3)

with no covariates used, or from the final hierarchical model containing covariates (Model ICC – see below for details).

Table 1: Minimum detectable effect size at different stages for the total sample and subgroups

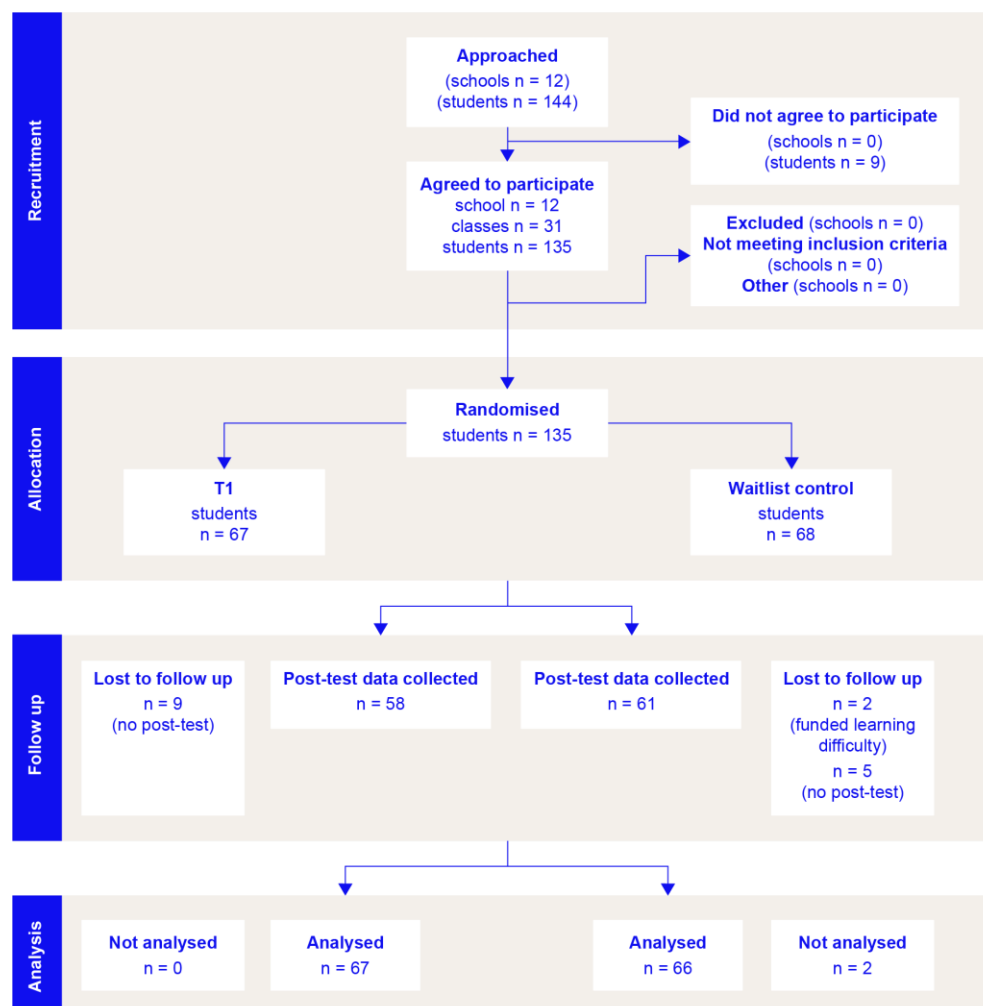
Stage	N [schools/ students] (n=intervention; n=control)	Correlation between pre- test (+other covariates) & post-test	ICC	Blocking/ stratificatio n or pair matching	Power	Alpha	Minimum detectable effect size (MDES) <i>f, d</i>
Total							
Protocol - expected	24/274 (137; 137)	0.5	0.05	School, 24 schools	80%	0.05	0.20
Randomisation (PAT-M Baseline)	23/288 (146; 142)	0.5	0.34	Class, 70 classes	80%	0.05	0.24
Analysis (PAT-M time 3)	23/257 (130; 127)	0.6	0.18	Class, 68 classes	80%	0.05	0.18
Analysis (model ICC)	23/257 (130; 127)	0.6	0.09	Class, 68 classes	80%	0.05	0.16
Primary							
Protocol - expected	12/137 (72; 72)	0.5	0.05	School, 12 schools	80%	0.05	0.33
Randomisation (PAT-M Baseline)	12/133 (67; 66)	0.5	0.17	Class, 30 classes	80%	0.05	0.38
Analysis (PAT-M time 3)	12/119 (58; 61)	0.6	0.16	Class, 30 classes	80%	0.05	0.28
Analysis (Model ICC)	12/119 (58; 61)	0.6	0.10	Class, 30 classes	80%	0.05	0.26
Secondary							
Protocol	12/137 (72; 72)	0.5	0.05	School, 12 schools	80%	0.05	0.33
Randomisation (PAT-M Baseline)	11/155 (79; 76)	0.5	0.07	Class, 40 classes	80%	0.05	0.24
Analysis	11/138	0.6	0.21	Class,	80%	0.05	0.26

(PAT-M time 3)	(72; 66)			38 classes			
Analysis (Model ICC)	11/138 (72; 66)	0.6	0.06	Class, 38 classes	80%	0.05	0.24

Follow-up

Participant flow is detailed in Figure 2. All of the suitable schools identified and approached by SCS consented to participate, with 94% and 97% of the invited students from the Primary and Secondary cohorts consenting to involvement respectively. Two control group students from the Primary cohort (<2%) and 14 students (6 intervention, 8 control) from the Secondary cohort (8%) were diagnosed with a learning disability during the trial period. These students continued with the QuickSmart intervention, but were excluded from all analyses as funded learning difficulties were listed as exclusion criteria for invitation into the study. These students have been removed from the participant flow diagram (Figure 2).

Figure 2: Participant flow diagram – Total cohort



Outcome measures

Assessment items were administered by the University of Newcastle researchers at all time-points. The researchers involved in administering and scoring the assessments were blinded to student allocation to intervention or control groups.

Primary outcome

Progressive Achievement Test – Mathematics (PAT-M)

The primary intervention outcome was measured with the Australian Council for Educational Research's Progressive Achievement Test – Mathematics (PAT-M) (ACER, 2018). The PAT-M is a rigorously tested measure of mathematics achievement that is well suited for evaluation of this project because each year level test is designed to be developmentally appropriate (e.g., Year 4; Year 8). Administered and scored using a paper-based format, participants completed the same level PAT-M at each assessment time-point (Year 4: PAT-M 3rd ed. Test Booklet 2; Year 8: PAT-M 3rd ed. Test Booklet 5).

The test scaled score was used for analysis. Scale scores are measures on an interval scale (0 to 100). This means that a difference of 5 in scale scores in the middle of the PAT scale (for example, from 50 to 55) is equivalent to the same difference on any other part of the scale (for example, from 15 to 20 or from 85 to 90). Scale scores allow comparison of results on test booklets of varying difficulty, and provide a common achievement scale for all test booklets.

Secondary outcomes

Student cognitive and affective measures

Mathematics self-beliefs have an impact on learning and academic achievement as they determine how well students are able to motivate themselves and persevere in the face of difficulties. They influence students' emotional life, and they affect the choices students make about their educational and career paths (Bandura, 1997; Wigfield and Eccles, 2000).

Self-beliefs were measured using instruments developed for the Programme for International Assessment (PISA). Administered in paper-based format, four of the PISA scales were used:

- Mathematics self-efficacy scale (MATHEFF), comprised of eight items.
- Interest in mathematics scale (INTMAT), consisting of four items.
- Self-concept scale (SCMAT), containing five items, and
- Mathematics anxiety scale (ANXIMAT), also containing five items.

Items were rated on a 4 point Likert scale ranging from (1) "Strongly agree" to (4) "Strongly disagree", with the average of the items used for analysis. The 2003 and 2012 PISA technical reports, and research from which the questions and calculation of measurement scales were based, highlight both valid and internally consistent scores (Wigfield et al., 1997; Ferla, Valcke & Cai, 2009).

Analysis

Primary analysis

The primary aim of the analysis was to assess whether the QuickSmart intervention had a significant impact on students' mathematics achievement, as measured by the 18-month post-intervention PAT-M test scores. A linear model predicting 18-month post intervention PAT-M scores was fitted. Baseline PAT-M scores were included as fixed effects to control for prior achievement. Gender and stage (Year 4 or Year 8) were included as fixed effects to adjust for these covariates, and group (intervention or control) was included as a fixed effect to examine if group allocation had a significant effect on PAT-M results. The regular mathematics class a student belonged to was included as a random intercept within the model to account for clustering of students within classes. Statistical analyses were completed using PASW Statistics 21 (SPSS Inc. Chicago, IL) software. Impacts were estimated using an intention-to-treat protocol. Alpha levels were set at $p < 0.05$. Group means and 95% confidence intervals (CIs) were determined using the linear model specified below.

For the i th student in the j th class, let Y_{ij} be the student's PAT outcome at 18 months. The model equation is

$$Y_{ij} = \beta_{0j} + \beta_1 x_{1ij} + \beta_2 x_{2ij} + \beta_3 x_{3ij} + \beta_4 x_{4ij} + r_{ij}$$

where $r_{ij} \sim N(0, \sigma^2)$ is random error and the intercept $\beta_{0j} = \gamma_{00} + u_{0j}$ consists of a random effect $u_{0j} \sim N(0, \tau_{00})$ of random variations around the overall mean γ_{00} . Binary predictors in the model are:

x_1 is Group allocation (Intervention or Control)

x_2 is Gender (Male or Female)

x_3 is Stage (Year 4 or Year 8)

and finally the model is adjusted for the student's baseline PAT score via the continuous covariate x_4 .

Note. This analysis was modified from that originally specified in the statistical analysis plan. The model design was modified from a linear mixed model, to a repeated measures ANCOVA model in order to align with the common analysis practice of EEF and SVA.

Interim analyses

Analysis of the effects of the QuickSmart intervention without a delayed post-test (Term 2, 2018) were undertaken using the immediate follow-up PAT-M results (Term 4, 2017). The

same linear model specified for the primary outcome analysis above, with the exception of let Y_{ij} as the student's PAT outcome at immediate follow-up.

This analysis was not specific in the original protocol.

Imbalance at baseline for analysed groups

There was a higher proportion of female participants among both cohorts (Table 2), however the balance by gender among randomised groups was marginal.

Table 2: Gender by cohort and randomised group

Cohort	% Female	% Male	N
Total			
Control	62.0%	38.0%	142
Intervention	60.2%	39.8	146
Total	61.0%	39%	288

Independent samples t-tests were used to evaluate whether random assignment resulted in equivalent groups at baseline for the primary and secondary outcomes (Table 3). There were no significant differences identified between groups. This result was the same among both cohorts.

Table 3: Baseline comparison of mathematics achievement (PAT-M scaled score)

	Intervention group		Control group		Effect size	
Outcome	N	Mean (SD)	N	Mean (SD)	Hedges g	p-value
Total	146	45.40 (7.28)	142	45.52 (7.39)	0.02	0.89
Primary	67	41.25 (6.54)	66	41.74 (7.11)	0.07	0.68
Secondary	79	48.92 (5.91)	76	48.79 (5.95)	0.02	0.88

Missing data

The extent, reasons for, and pattern (potential mechanism for missingness) were investigated for the primary outcome variable (PAT-M scaled score). The missingness procedure was triggered if more than 10% of data for a single variable, or single class was

missing. The first step was to assess whether the data was missing at random (MAR) using a logistic regression to test whether missing cases could be predicted using the covariates in primary analysis model (Sex, Stage number, group) and the baseline PAT-M scaled score. Where predictability was confirmed, multiple imputation (MI) was undertaken. All variables in the predictive logistic regression were to be used in an MI strategy using a fully conditional specification within SPSS 24 MI to create 10 imputed data sets. Treatment effects were to be re-estimated using each dataset, with Rubin combination rules used to find the average and estimate standard error. Assessment of the sensitivity of the estimate to missingness was via comparison of the complete data to the imputed estimates. If the complete data only model confirmed effectiveness but the imputed estimate did not, we assumed that the missing data was missing not at random to such an extent as to invalidate our conclusion of effectiveness.

Where the missingness could not be predicted, we assumed the data was either 'Missing Completely at Random' (MCAR) or 'Missing Not at Random' (MNAR). Reasons for missing data (e.g., student absence on day of testing, student withdrawn from school) were considered to support the argument for classification as MCAR, and MI was not considered feasible. If data was considered to be MNAR, a structural modelling approach would be the only option, and was not feasible as it would deviate from the principles of transparent reporting due to being assumption rather than data driven.

Overall, 10.4% ($n = 30/288$) of cases for the PAT-M scaled score were lost to follow-up. The logistic regression procedure described for determining the mechanism of missingness was not significant, indicating the mechanism was not MAR. Given that 21/30 (7.1%) and 5/30 (1.7%) missing cases at follow-up were due to students moving school and sickness on the day of testing respectively, with only 4/30 (1.4%) missing cases due to study withdrawal, the missing data was assumed to be MCAR.

Secondary outcome analyses

Secondary outcomes were assessed using an intention-to-treat protocol. The model detailed in the primary analysis was applied to the secondary outcomes.

Treatment effects in the presence of non-compliance

Fidelity checklists from each instructor detail the number of sessions undertaken by each student within the intervention group across the intervention period (90 sessions expected for 100% dosage). Frequencies of the program dosage quartiles are reported.

In a Randomised Controlled Trial the intervention effect in an intention-to-treat (ITT) analysis may be biased when there is 'contamination' by participants who received a different level of the intervention protocol than their random allocation prescribed (e.g., for reasons such as non-compliance, partial compliance, or issues with protocol delivery). An Instrumental Variables (IV) approach (Angrist & Imbens, 1995) may be used to estimate the unbiased treatment effect by using the received treatment as an instrument for group allocation. This

approach tends to be more rigorous than per-protocol or on-treatment analyses (McNamee, 2009; Tillbrook et al. 2014).

A Two Stage Least Square (2SLS) approach was undertaken using the SYSLIN procedure in SAS. The first stage of the 2SLS involved regression of group allocation (intervention = 1; control = 0) on the continuous compliance instrument (proportion of QuickSmart sessions undertaken), with covariates: baseline PAT-M scaled score, gender and stage number. The second stage involved regression of the dependent variable (Follow-up PAT-M scaled score) on the predicted values obtained from the first stage, with covariates baseline PAT-M scaled score, gender and stage number.

The correlation between group allocation and compliance variables is reported alongside the model F test from the first stage. The suggested rule of thumb for checking the strength of the instrument is that the F-statistic from the first stage regression should be greater than 10, or the t value of the instrument above approximately 3 (Angrist, 2006), otherwise the instrument is considered to be weak, the consequence of which is that the sampling distribution of the 2SLS estimator might not be approximately normal even in large samples.

The parameter estimates and corresponding effect size from the second stage are reported. It is also important to note that the standard errors produced in the second stage regression are not correct; the standard error for the 2SLS estimates must take into account the additional uncertainty due to performing two stages of regression. The SYSLIN procedure in SAS contains an algorithm to adjust the standard errors in the second stage.

This analysis was not pre-specified in the original trial protocol.

Subgroup analyses

Primary and Secondary cohorts were investigated as sub-groups for analysis of primary outcomes. In each case, the statistical procedure for the analysis of the whole group analysis were followed, with the exception of the removal of the stage (Year 4 or Year 8) covariate from models.

Effect size calculation

Hedges' g was used to determine effect sizes of the change in mean score for each group relative to the baseline value (effect of intervention on the change score).

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

where our conditional estimate of $\bar{x}_1 - \bar{x}_2$ is recovered from β_1 in the primary ITT analysis model;

s^* is estimated from the analysis sample as follows:

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

where n_1 is the sample size in the control group, n_2 is the sample size in the treatment group, s_1 is the standard deviation of the control group, and s_2 is the standard deviation of the treatment group (all estimates of standard deviation used are unconditional).

Ninety-five per cent confidence intervals (95% CIs) of the effect size were computed using the `compute.es` function in R. This function computes the confidence intervals using the variance in g derived by the Hedges & Olkin (Hedges & Olkin, 1985, p. 86) formula:

$$var(g) = \frac{n_1 + n_2}{n_1 n_2} + \frac{g^2}{2(n_1 + n_2)}$$

References:

- Angrist, J. (2006). Instrumental variables methods in experimental criminological research: What, why and how. *Journal of Experimental Criminology*, 2, 23–44.
<https://doi.org/10.1007/s11292-005-5126-x>
- Angrist, J., & Imbens, G. (1995). Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *American Statistical Association*, 90(430), 431–442.
- Baker, S., Gersten, R., & Lee, D.S. (2002). A synthesis of empirical research on teaching mathematics to low-achieving students. *Elementary School Journal*, 10, 51–73.
- Deshler, D., Mellard, D. F., Tollefson, J. M., & Byrd, S. E. (2005). Research topics in responsiveness to intervention. *Journal of Learning Disabilities*, 38(6), 483–484.
- Fuchs, D. & Fuchs, L. S. (2001). Principles for the prevention and intervention of mathematical difficulties. *Learning Disabilities Research and Practice*, 16, 85–95.
- Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.
- Marston, D. (2005). Tiers of intervention in responsiveness to intervention: Prevention outcomes and learning disabilities patterns. *Journal of Learning Disabilities*, 38(6), 539–544.
- McMaster, K. L., Fuchs, D., Fuchs, L. S., & Compton, D. L. (2005). Responding to nonresponders: An experimental field trial of identification and intervention methods. *Exceptional Children*, 71(4), 445–463.
- McNamee, R. (2009). Intention to treat, per protocol, as treated and instrumental variable estimators given non-compliance and effect heterogeneity. *Statistics in Medicine*, 28(21), 2639–2652.
- Rowe, K., Stephanou, A., & Urbach, D. (2006). Effective teaching and learning practices initiative for students with learning difficulties. Report to the Australian Government Department of Education, Science and Training (DEST). Canberra: Australian Government Printery.
- Tilbrook, H. E., Hewitt, C. E., Aplin, J. D., Semlyen, A., Trehwela, A., Watt, I., & Torgerson, D. J. (2014). Compliance effects in a randomised controlled trial of yoga for chronic low back pain: a methodological study. *Physiotherapy*, 100(3), 256–262.
- Westwood, P. (2007). *What teachers need to know about numeracy*. Melbourne: ACER.

- Bloom, H. S., Richburg-Hayes, L., & Black, A. R. (2007). Using Covariates to Improve Precision for Studies That Randomize Schools to Evaluate Educational Interventions. *Educational Evaluation and Policy Analysis*, 29(1), 30-59. doi:10.3102/0162373707299550
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Earlbaum Associates.
- Cohen, J. (1992). A power primer. *Quantitative methods in psychology*, 112(1), 155-159.
- Donner, A., & Klar, N. (2000). *Design and Analysis of Cluster Randomization Trials in Health Research*: Arnold.
- Dreyhaupt, J., Mayer, B., Keis, O., Öchsner, W., & Muche, R. (2017). Cluster-randomized Studies in Educational Research: Principles and Methodological Aspects. *GMS journal for medical education*, 34(2), Doc26-Doc26. doi:10.3205/zma001103
- Hedges, L., & Olkin, I. (1985). *Statistical Methods for Meta-Analysis*. New York: Academic Press.
- Lamb, S., & Fullarton, S. (2001) Classroom And School Factors Affecting Mathematics Achievement: a Comparative Study of the US and Australia Using TIMSS. In. http://research.acer.edu.au/timss_monographs/10.