Getting it right from the start: An early years' school intervention to equitably improve student oral language and reading outcomes

Statistical Analysis Plan

Version 1.0
04Apr2024

HREC #: 2019.1 77

Author: Francesca Orsini
Clinical Epidemiology & Biostatistics (CEBU)
Melbourne Children's Trial Centre (MCTC)
Murdoch Children's Research Institute
The Royal Children's Hospital, 50 Flemington Road
Parkville, Victoria 3052 Australia

CONTENTS

List of Abbreviations

AEDC	Australian Early Development Census
CAC	Cluster autocorrelation
CELF-5	Clinical Evaluation of Language Fundamentals – 5th Edition
CEM	Catholic Education Melbourne
CI	Chief Investigator
CI	Confidence Interval
DE	Department of Education
EBF	Eureka Benevolent Foundation
GIRFTS	Getting it Right from the Start
HREC	Human Research Ethics Committee
ICC	intra cluster correlation coefficient
ICSEA	Index of Community Socio Educational Advantage
ISP	Implementation Support Partner
ITT	Intention to Treat
IQR	Interquartile Range
MACS	Melbourne Archdiocese Catholic Schools
MCRI	Murdoch Children's Research Institute
mITT	modified intention to treat
PL	Professional Learning
RCT	Randomised Controlled Trial
RPT	Reading Progress Test
RTI	Response to Intervention
SAP	Statistical Analysis Plan
SD	Standard Deviation
SOLAR	Science of Language and Reading
SPAT	Sutherland Phonological Awareness Test
SW CRT	stepped wedge cluster randomised controlled trial
TOWRE2	Test of Word Reading Efficiency – 2nd Edition
UoM	University of Melbourne

1. ADMINISTRATIVE INFORMATION

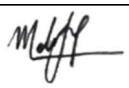
Protocol: HREC #: 2019.177; Version 17, 25 October 2023

1.1 Document Version History

Version Date	Version	Author	Signature	Change Description	Reason/Comment
04 April 2024	1.0	Francesca Orsini		Initial release.	Not applicable.

1.2 Approvals

The undersigned have reviewed this plan and approve it as final. They find it to be consistent with the requirements of the protocol as it applies to their respective areas. They also find it to be compliant with ICH-E9 principles and confirm that this analysis plan was developed in a completely blinded manner (i.e., without knowledge of the effect of the intervention being assessed)

Name	Role on Study	Affiliation	Signature	Date
Prof Sharon Goldfeld	Principal Investigator	MCRI, RCH		11/4/24
A/prof Jon Quach	Co-investigator	MCRI, UoM		10/4/2024
Ms Beth Shingles	Co-investigator	MCRI, UoM		09.04.2024
Ms Melissa Cheah	Project team	MCRI		8/4/2024
Ms Cecilia Sinclair	Project team	MCRI		8/4/24

2. STUDY SYNOPSIS

Early intervention in oral language and reading skills has the potential to make a difference for children's development outcomes. Exemplary teaching and effective and timely intervention can lead to high levels of achievement for 'at risk' or vulnerable students (Buckingham et al., 2013; Snow et al., 2014). The first years of school are a key timepoint to effect this change through evidence-informed interventions, establishing solid foundational skills that enable students to learn in subsequent years (Gillon et al., 2022). Given this, there is an opportunity to make a substantial difference to educational and life outcomes by addressing gaps in language abilities and early reading skills. The challenges for many schools and teachers are (1) identifying those children who are at risk; (2) identifying evidence-based interventions that address student needs; and (3) effectively implementing interventions.

The Response to Intervention (RTI) framework is a promising model to address these challenges, meeting the needs of all learners and addressing current inequalities. This preventative model proactively aims to help children 'keep up', rather than waiting for them to need to 'catch up'. The components of RTI are: (a) identification of students needing additional support through screening; (b) student data analysis informing the selection of appropriate interventions; (c) multi-tiered supports dependent on student needs; and (d) progress monitoring to measure impact (Fletcher & Vaughn, 2009; Bradley et al., 2005). When implemented with fidelity, RTI has been shown to lead to improved student outcomes (Bianco, 2010). It is estimated that the incidence of reading difficulties can be dramatically reduced where high quality whole-of-class teaching is supplemented with targeted small groups for 'at risk' learners (Gilbert et al., 2013).

There is growing research regarding the implementation of RTI in the United States, where the approach is widespread (Noell & Gansle, 2016). School-level factors such as resources, timing, expertise and time have been found to considerably impact a school's ability to implement RTI with fidelity (Lopuch, 2018). Similarly, differences in implementation success have been attributed to a variety of factors, including educator buy-in and participation, and attention to the organisational context (Redding, et al., 2017). Without effective implementation support strategies, evidence-based interventions may not lead to sustainable change (Curran et al., 2012; Goldfeld et al., 2022b). Ongoing support strategies such as coaching of teachers have contributed to improvements in outcomes and are also critical for sustaining a practice over time (Freeman et al., 2017; Horner et al., 2017). Implementation 'systems' coaching is a professional development approach where coaches help teachers and schools build the implementation fidelity of evidence-based practices and whole-school reforms such as RTI (March et al., 2020).

The Getting it Right from the Start (GIRFTS) is a stepped-wedge cluster randomised controlled trial (SW-CRT) of a RTI model in the first two years of school (Foundation and Grade 1) to improve oral language and reading student outcomes at the end of Grade 1 when compared to business as usual. Clusters are defined as the schools. Each period is one Australian school year (end January – December, 4 terms in total, 9-11 weeks per term). The study design consists of three periods (period 0,1,2) and two cohorts. Up to 18 schools were to be randomised into one of two cohorts by an independent statistician. The first cohort of schools stepped into the intervention in period 1 (2022), whilst schools in cohort 2 crossed over from control ('business as usual') into the intervention in period 2 (2023), as reported in Table 1.

Table 1: Trial design

<table>
<tr><th colspan="4">2021
period 0</th><th colspan="4">2022
period 1</th><th colspan="4">2023 – period 2</th><th colspan="2">2024</th></tr>
<tr><th>T1</th><th>T2</th><th>T3</th><th>T4</th><th>T1</th><th>T2</th><th>T3</th><th>T4</th><th>T1</th><th>T2</th><th>T3</th><th>T4</th><th>T1</th><th>T2-4</th></tr>
<tr><td colspan="3" rowspan="2">Business as usual
Foundation & G1</td><td rowspan="2">Randomisation</td><td>Cohort 1</td><td>data collection (G2)</td><td colspan="2">RTI
Foundation & G1</td><td>data collection (G2)</td><td colspan="3" rowspan="2">RTI
Foundation & G1</td><td rowspan="2">data collection (G2)</td><td rowspan="2">Analysis and write up</td></tr>
<tr><td>Cohort 2</td><td>data collection (G2)</td><td colspan="2">Business as usual
Foundation & G1</td><td>data collection (G2)</td></tr>
</table>

2.1 Objectives

The primary objective of the trial is to determine the impact of the RTI model (Tier 1 and Tier 2, as explained in section 2.4), compared with 'business as usual', at the end of Grade 1 on students' oral language and reading outcomes.
The secondary objective of the trial is to gain an in-depth understanding of the process of RTI implementation at the teacher, school, and sector-level by investigating factors that both promote and inhibit the success of this approach.
We hypothesise that compared with 'business as usual', RTI will result in improved students oral language and reading outcomes.

The analysis of secondary objective will be covered by separate analytic plans. The current SAP details the analysis of the students outcomes to answer the primary objective.

2.3 Study Population

2.3.1 Schools (cluster eligibility)

To be included in the study, schools had to:
- have language and cognitive domain vulnerability rates of ≥ 10% (2018 AEDC results)
- have their Index of Community Socio-Educational Advantage (ICSEA) value ≤ 1100
- be located in North-East or North-West region of Department of Education OR be located in the Melbourne Archdiocese Catholic Schools (MACS)

2.3.2 Teachers and school staff during the Intervention Period

All teachers in participating schools during the intervention period were included who:
- Signed a Consent Form
- Were in Foundation or Grade 1 classes between 2022 and 2023 (Cohort 1) and 2023 (Cohort 2)
- Were involved in teaching Tier 2 small-groups (this includes Education Support staff) in Foundation or Grade 1 classes between 2022 and 2023 (Cohort 1) and 2023 (Cohort 2)
- Were involved in Literacy Leadership in the school

Principals and Assistant/Deputy Principals who signed a consent form were also involved in completing surveys (SOLAR and RTI screener)

2.3.3 Students

- All Foundation and Grade 1 students (5-7 years old) enrolled in participating schools between 2021 and 2023.
- Students whose parents/guardians agreed to participate in the study (recruited through an active consent process for baseline data collection and subsequently an opt-out process)

2.4 Intervention

The RTI model implemented in this study is described as two tiers (Tier 1 and Tier 2), offering increasing levels of support. Tier 1 represents high-quality evidence-based whole-of-classroom instruction provided to all students. Tier 2 is more intensive teaching support, often provided in small groups targeting specific skills. These tiers are used to describe the type, format, and setting of learning support students receive relative to their peers. The principles that drive RTI and make it a dynamic and responsive framework to improve student outcomes are: (a) early identification of students needing additional support through screening; (b) student data analysis informing the selection of appropriate interventions; (c) multi-tiered supports dependent on student needs; and (d) progress monitoring to measure impact.

2.5 Randomisation and Blinding

Schools were randomly assigned to either cohort 1 or cohort 2. Cohort 1 commenced the intervention implementation in 2022 (period 1), while cohort 2 was still conducting 'business as usual', and then commenced the intervention in 2023 (period 2).
A statistician independent from the trial prepared the randomisation schedule using block randomisation to maintain balance between cohorts. Randomisation was stratified by school sector and DE region with variable block sizes (1-CEM schools, 2-Victorian DE North East region schools, 3-Victorian DE North West region schools). Schools to be randomised were ordered alphabetically by school name within each education sector.

The study coordinator, research assistants, implementation support coaches, randomising statistician and teachers were aware of the allocation to either cohort to enable study organisation and implementation.
Staff responsible for conducting the oral language and reading assessments at all time points were blinded to randomisation allocation. School staff and teachers were asked not to disclose their randomisation status during this assessment.

2.6 Sample Size

Existing trials in schools show relatively modest effects (effect sizes of 0.2–0.4 SDs) for outcomes such as child literacy and language development from the early years of primary school. While effect sizes of 0.20–0.3 SDs can be meaningful and impactful at the whole population level, targeted primary school interventions such as GIRFTS involve a cost and intensity such that larger effects in the short-to-medium term might be necessary to justify implementation at a population level. Given the primary objectives of the trial and measures collected in existing RCTs, we chose to anchor our sample size calculation around the detection of a minimum effect size of 0.28 SD in reading comprehension measured via the Reading Progress Test (RPT) to allow comparisons with other international programs.
Sample size calculation was performed for the effect of the intervention on the primary outcome (RPT at the start of Grade 2) using "The Shiny CRT Calculator: Power and Sample size for Cluster Randomised Trials" (https://clusterrcts.shinyapps.io/rshinyapp/). We assumed the following: a standard deviation of 1, an intra-cluster correlation coefficient (ICC) of 0.02, a Cluster autocorrelation (CAC) of 0.8, and an alpha of 0.05. A total of 16 schools (8 schools in each cohort) and an average number of 35 students per year per school (who give consent and have primary outcome data available) were required to detect an effect size of 0.28 SD on the RPT score between intervention and control groups with a power of 80%. To account for the potential dropout of up to 2 schools during the study, we recruited 18 schools (9 in each cohort).
If none of the 18 schools drop out from the study, an average of 30 students per year per school would be adequate to detect a difference of 0.28 SD on the RPT score between groups with a power of 80%.

2.7 Deviations to original protocol due to the COVID-19 Pandemic related issues

In the early phase of the study, within one-month post-randomisation, three schools expressed willingness to withdraw from participation into the study before the commencement of period 0 collection (2021). This was due to changes in staffing impacted by COVID-19 and competing priorities. To mitigate potential loss of statistical power and ensure the fulfillment of predetermined sample size requirements (at least 16 schools as per section 2.6), a strategic decision was made to conduct a second round of randomisation. This approach aimed to safeguard against the potential reduction of the sample size below the stipulated threshold of 16 schools. Consequently, an additional school was successfully recruited, contributing to the preservation of the intended study sample and the maintenance of statistical robustness.

2.8 Outcomes

Table 2: Outcomes

Outcome	Outcome Measure	Outcome Measure Details
Student oral language	The Clinical Evaluation of Language Fundamentals **CELF** – Australian and New Zealand standardised Fifth Edition Subtest: Following Directions	- A research-based, individually administered clinical tool designed to assess students morphology, syntax, semantic skills, and pragmatics. - For ages 5 years 0 months through to 21 years 11 months - Can be administered in approximately 15 minutes (5 mins per subtest) - Administered by blinded data collector volunteers in at least the third year of a relevant tertiary qualification or undertaking a masters degree (trained by a registered Speech Language Pathologist and supervised by blinded research assistants) - administered to all grade 2 students at all participating schools (i.e., Cohort 1 and 2) at the beginning of each year of the study
	CUBED: Narrative Language Measures (NLM) listening subtest	- Assesses a student's ability to comprehend and produce complex, academic language. - Students asked to listen to a brief narrative, and then retell that narrative. This incorporates listening comprehension, cognition, memory, and expressive language skills. - Scoring takes place in real-time. - Additional sections of the NLM allow for the examination of story grammar comprehension and inferential word learning or word use. - Administered by blinded data collector volunteers in at least the third year of a relevant tertiary qualification or undertaking a masters degree (trained by a registered Speech Language Pathologist and supervised by blinded research assistants) - administered to all consented grade 2 students at all participating schools (i.e., Cohort 1 and 2) at the beginning of each year of the study
Student reading	Student reading The Reading Progress Test (**RPT**) [primary outcome]	- Assesses pre-reading and early reading skills, including phonological awareness, print concepts, word knowledge, and comprehension via administration to an individual or group of students, yielding standard scores. - validated tool and has Australian norms. - Administered by blinded data collector volunteers in at least the third year of a relevant tertiary qualification or undertaking a masters degree (trained by a registered Speech Language Pathologist and supervised by blinded research assistants) - Administered as a group test but provides an individual score of reading ability per child. - Administered to all consented grade 2 students at all participating schools (i.e., Cohort 1 and 2) at the beginning of each year of the study.
	Test of Word Reading Efficiency–Second Edition (**TOWRE 2**) – Form B	- A measure of an individual's ability to pronounce printed words (Sight Word Efficiency) and phonemically regular non-words (Phonemic Decoding Efficiency) accurately and fluently. The TOWRE 2 provides an efficient means of monitoring the growth of two

Outcome	Outcome Measure	Outcome Measure Details
		kinds of word reading skill that are critical in the development of overall reading ability. - The Sight Word Efficiency (SWE) subtest assesses the number of real words printed in vertical lists that an individual can accurately identify within 45 seconds. - The Phonemic Decoding Efficiency (PDE) subtest measures the number of pronounceable non-words presented in vertical lists that an individual can accurately decode within 45 seconds. The four forms of each subtest are of equivalent difficulty, and any of the forms of each subtest may be given depending on the purposes of the assessment. If only one form of each test is used, the test can be administered in approximately 5 minutes, including time for directions and practice items. - Administered by blinded data collector volunteers in at least the third year of a relevant tertiary qualification or undertaking a masters degree (trained by a registered Speech Language Pathologist and supervised by blinded research assistants) - administered to all consented grade 2 students at all participating schools (i.e., Cohort 1 and 2) at the beginning of each year of the study.
	Sutherland Phonological Awareness Test (SPAT-R)	- Individually administered test of phonological and phonemic awareness - Assesses identification and manipulation of syllables, rhymes and phonemes, and includes tests of non-word reading and spelling - Norm referenced - Takes 10-15 minutes to administer - Administered by blinded data collector volunteers in at least the third year of a relevant tertiary qualification or undertaking a masters degree (trained by a registered Speech Language Pathologist and supervised by blinded research assistants) - It will be administered to all consented grade 2 students at all participating schools (i.e., Cohort 1 and 2) at the beginning of each year of the study

Full details of the background to the trial and its design are presented in the protocol.

3. GENERAL STATISTICAL METHODOLOGY

3.1. Analysis Software

All analyses will be performed using Stata Release 18 or later.

3.2. Definition of Control and Intervention data periods

As shown in Table 1, in year 2021 (period 0) both cohorts were in the control period; data collection for this period occurred on Grade 2 students at the start of year 2022, e.g. during term 1. Randomisation of schools into cohorts occurred at the end of the school year 2021 (period 0), before period 0 data collection started.
Cohort 1 schools started RTI implementation at the beginning of 2022 (period 1) whilst for cohort 2 schools RTI implementation started at the beginning of 2023 (period 2): in both years there was some overlap between data collection (during term 1) and beginning of the teachers professional learning, which is the first implementation activity. We acknowledge some risk of contamination of control data for these periods, but this was deemed acceptable for the following reasons. The majority of student data collection was expected to be completed before the conclusion of professional learning. The RTI intervention was delivered to Foundation and Grade 1 teachers and students, whereas outcome data was collected from students in Grade 2. It is expected that little contamination would occur between year levels because very few Grade 2 teachers would attend GIRFTS professional learning. If Grade 2 teachers did attend GIRFTS professional learning, it was expected that there would be insufficient time for them to adjust their teaching practice before Grade 2 student assessments are complete. The majority of Grade 2 data collection was expected to be complete in Term 1 with follow up of any remaining students in Term 2.

3.3. Definition of Analysis Populations

The primary population of interest is the modified intention-to-treat (mITT) population, which includes all randomised schools who undergo period 0 data collection, with data analysed according to the intervention that was supposed to be delivered in that particular period, irrespective of whether the intervention had been implemented as planned.

3.4. Adjustment for Multiplicity

No formal adjustments for multiplicity of testing will be applied, with results interpreted based on the magnitude of the estimates and the 95% confidence intervals (CIs) rather than focussing on p-values.

3.5. Interim Analyses

No interim analyses were/will be conducted.

3.6. Handling of Missing Data

Outcome data could be missing in different situations:

- where schools withdrew from the study (generally monotone missingness). It is reasonable to assume that when schools withdrew from the study data is missing at random (MAR).
- where students were absent during data collection period, or students moved school prior to data collection period, or parents withdrew consent to data collection. Even under these circumstances, it is reasonable to assume that the mechanism that generates missing data is MAR.

If the proportion of missing outcome data is ≤10% in the primary outcome, we will assume that data is MAR and an analysis by maximum likelihood will produce an unbiased estimate of the intervention (note: the sample size was inflated by 2 schools to allow for loss of power due to monotone missing data).

However, if the proportion of missing outcome data is > 10% in the primary outcome, missing data in the primary and secondary outcomes will be handled using multiple imputation. Multiple imputation will be conducted using chained equations. Ideally this would be conducted using a single imputation model for all outcomes, although it may be necessary to impute outcomes using separate models for each outcome. Ordinal variables will be imputed using ordinal regression and continuous variables will be imputed using linear regression, or predictive mean matching if non-normal. Baseline variables will be included as auxiliary variables in the imputation model. Imputation will be carried out using 50 imputed datasets. The multilevel structure of the data will be taken into account in the imputation model.

4. DESCRIPTIVE STATISTICS

4.1. Recruitment and Follow-up of Clusters

All schools (clusters) who were invited to participate in the GIRFTS trial will be accounted for as part of the CONSORT flow diagram.

4.2. Cluster Characteristics

Schools randomised will be described by:

- school sector/DE region (North-East of DE / North-West region of DE / MACS ([N(%)]
- Number of teachers who implemented RTI
- Number of leaders who implemented RTI
- Number of Grade 1 students participating in the trial [Median (Range)]
- Language and cognitive domain vulnerability rates [Median (Range)]
- ICSEA [Median (Range)]

4.3. Students

The characteristics of students will be summarised separately by those in the control and the intervention period, across all schools. The following data will be summarised:

- Gender: Male/Female/Null, Not Recorded, Other [N(%)]
- Age at assessment [mean(SD)]

We will also present a plot of the cluster size (number of students) over time (years) indicating control and intervention periods using different symbols/colours.

At the time of writing this SAP, the study team was discussing the option to do data linkage with school census data in order to collect additional details on the students involved in the study, including but not limited to: country of birth, support through the Tutor Learning Initiative, Aboriginal and Torres Strait Islander status, presence of disability, Language Background Other than English (LBOTE), Student Family Occupation (SFO), Student Family Education (SFE). Should data linkage be successful, this data also will be summarised separately by those in the control and the intervention period, across all schools.

4.4. Engagement with the Intervention

Engagement with the RTI model will be summarised during the intervention period by the following characteristics, for all clusters by year of the intervention:

- School/teacher/leaders attendance at the training
- School/teacher/leaders attendance at the community of practice sessions
- School/teacher/leaders attendance at reflection and planning meetings
- N weeks of the year that ISP went to the school

Non-engagement with the intervention could be due to school-level factors (e.g., the school stops being engaged with the study team) or teachers-level factors (e.g., teachers not attending or partially attending training sessions) or students-level (e.g., students miss school, move school). For analysis purposes, we will take into account no engagement from a school level only, and non-engagement will be defined as:

- No teachers sent to Professional Learning sessions AND
- No attendance at reflection and planning meetings AND
- ISP attendance at school less than 50%

5. ANALYSIS OF PRIMARY AND SECONDARY OUTCOMES

5.1 Primary Outcome: Student reading The Reading Progress Test

Description: The Reading Progress Test (RPT) is a validated tool with Australian norms which assesses pre-reading and early reading skills, including phonological awareness, print concepts, word knowledge, and comprehension via administration to an individual or group of students, yielding raw scores and standard scores. A higher score indicates better comprehension of written text. RPT standard scores are generated from raw scores based on student's grade (see Stata syntax in APPENDIX B).

- RPT raw score - range 0-32.
- RPT standard score – range 70-130. The table below reports the conversions from raw scores to standard scores for grade 2 students.

Table 3: conversion from RPT raw score to RPT standard score

Raw Score	0/5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
Standard Score	70	71	73	75	76	78	79	80	81	82	83	84	85	86	88	89	91	92	94	96	98	100	103	106	110	117	124	130

- Categorisation of RPT standard score
  - 70/73 =1 "well below average"
  - 74/88 =2 "below average"
  - 89/111 =3 "average"
  - 112/126 =4 "above average"
  - 127/130 =5 "well above average"

Target population: Foundation and Grade 1 Students enrolled in Public and Catholic schools with language and cognitive domain vulnerability rates of ≥ 10% and ICSEA ≤1100 in inner city and regional Victoria.

Treatment conditions: RTI model (intervention period) vs business as usual (control period)

Endpoint: a. RPT standard score at the beginning of Grade 2
b. RPT raw score at the beginning of Grade 2
c. RPT category at the beginning of Grade 2 (only descriptive statistics)

Population-level summary measure: Between group (intervention vs. control) adjusted mean difference.

Possible Intercurrent Events:

Event	Handling of Event	Justification
Non-engagement with the intervention (as defined in section 4.5)	Treatment Policy	We would like to estimate the effect of the intervention regardless of the intercurrent event, thus reflected a real-world estimate of the model on referral patterns. Outcome data (RPT) will be used in the analysis, if available, irrespective of the event e.g., irrespective of whether the school engaged in the RTI model.

Primary Analysis:
The primary outcome (RPT) will be analysed using mixed effects model with a gaussian distribution with an identity link function, fitted to data at the students level. The model will include fixed effects for:

- group (intervention vs control),

- calendar time (year),
- school sector/DE region (MACS schools vs Victorian DE North East region vs Victorian DE North West region)

and random effects for:

- school (cluster),
- grade 1 class/teacher within school.

We assume an exchangeable within-cluster correlation structure with a single cluster (school) random effect (e.g., correlation between any two students in the same cluster is constant). Deviations from this will be considered as sensitivity analyses.
We will report a model-fitted marginal mean difference (students-average treatment effect (1)) and its 95% confidence interval (CI) estimated using the margins command in STATA. We will plot means and SDs of RPT over time (years) indicating control and intervention periods using different symbols/colours.

If computational difficulties arise when fitting the analysis model, we will take alternative analytical approaches as follows:

1. Without school sector/DE region fixed effect
2. Without grade 1 class/teacher-level random effect
3. Without school-level random effect

Missing Data: refer to section 3.6.

Sensitivity Analyses:

1) Our primary model includes a fixed effect for time, implying a common underlying secular trend across all schools. We will consider whether there is a different secular trend for stratum defined by geographical location (North East DE / North West DE / MACS)) through including a fixed-effect interaction between time and location stratum in the model. If an interaction effect does present itself (in terms of a significant interaction and a substantial change in the conclusions of the intervention effect), we will adjust for the interaction effect in our primary analysis.
2) Our primary model includes a single fixed effect for intervention. This assumes that the effect of the intervention is the same for all schools. We will examine whether the effect of the intervention varies by school by incorporated a fixed-effect interaction between group (intervention vs control) and school (cluster). As our primary interest is in the average effect of the intervention this will be reported in addition to the primary analysis if found to be significant.
3) We will check for evidence of informative cluster size (i.e., the intervention effect differs between smaller and larger clusters). If there is evidence of informative cluster size, we will reanalyse using cluster-robust standard errors.
4) We will check for evidence of correlation structure misspecification. We assume an exchangeable within-cluster correlation structure in the main analysis with a single cluster (school) random effect (e.g. correlation between any two students in the same cluster is constant). Will undertake a sensitivity analysis assuming a block-exchangeable correlation structure (e.g. correlation between two students in the same period is higher with lower level of correlation for students in different periods). This will be done by extending the random-effects components to allow a random interaction between time and school.

5) We will reanalyse data including adjustment for baseline factors associated with the primary outcomes, including (in order of importance, inclusion dependent of ability of model to converge): student gender and student age at assessment.
6) [if data linkage be successful, we will further analyse data including adjustment for baseline factors associated with the primary outcomes, including (inclusion dependent of ability of model to converge: student gender, student age at assessment, country of birth, support through the Tutor Learning Initiative, Aboriginal and Torres Strait Islander status, presence of disability, Language Background Other than English, Student Family Occupation (SFO), Student Family Education (SFE)].

Supplementary Analyses:
We will perform a supplementary analysis aimed at estimating the effect of the intervention in a subset of students enrolled in schools who would engage with the RTI model if offered it (i.e., Compliance Average Causal Effect -CACE). In order to do this analysis, we will split our population into four principal strata: known as compliers (for whom the intervention received is the same as the intervention assigned), never-takers (who do not receive intervention, regardless of intervention assignment), always takers (who receive intervention regardless of intervention assignment), and defiers (who always do the opposite of their intervention assignment). The RTI model is controlled by investigators and is not available to schools during the control period, as such, always-takers and defiers cannot exist, reducing the possible strata to compliers and never-takers.
There is no a priori definition of the level of engagement required which we can use to inform our definition of compliers. Therefore, for this analysis, we will define compliers as those schools who engage at all, in opposition to those schools who do not engage, as per definition of non-engagement with the intervention reported in section 4.5. We will then conduct a subgroup analysis to attain the treatment estimate in 'compliers' (defined at each level of engagement discussed above) using methods detailed by Gruber et al. (2).

5.2 Secondary Outcome: Clinical Evaluation of Language Fundamentals (CELF)

Description: Following Directions is one subtest of the Clinical Evaluation of Language Fundamentals®–fifth Edition (CELF®–5). The CELF is an individually administered test for determining if a student (ages 5 through 21 years) has a language disorder or delay. Following Directions takes 7-10 minutes to administer and sees the student point to pictured objects in response to oral directions (Pearson, 2013). In this study, CELF-5 Following Directions subtest will be collected at the beginning of Grade 2.
Derived scaled scores will be converted from raw scores using the CELF manual, and have a mean 10 and standard deviation of 3. Scaled scores are used to compare students' performance to typical performances of the same-age norm group. These scores are derived from the total raw scores for each test and are on a normalized score scale. Score tables are derived from a normative sample (Pearson, 2013 - see Stata syntax in APPENDIX C).

Target population: Foundation and Grade 1 Students enrolled in Public and Catholic schools with language and cognitive domain vulnerability rates of ≥ 10% and ICSEA ≤1100 in inner city and regional Victoria.

Treatment conditions: RTI model (intervention period) vs business as usual (control period)

Endpoint: CELF Following Directions subscale at the beginning of Grade 2

Population-level summary measure: Between group (intervention vs. control) adjusted mean difference

Possible Intercurrent Events: As described for the primary outcome (see Section 5.1).

Analysis: The analysis will be similar to the one of the primary outcome (see Section 5.1).

Supplementary Analyses: The analysis will be similar to the one of the primary outcome (see Section 5.1).

5.3 Secondary Outcome: CUBED: Narrative Language Measures (NLM) Listening Retell Sub-test (LRS)

Description: CUBED assesses a student's ability to comprehend and produce complex, academic language. Students are asked to listen to a brief narrative, and then retell that narrative. This incorporates listening comprehension, cognition, memory, and expressive language skills. Scoring takes place in real-time, whilst additional sections of the NLM allow for the examination of story grammar comprehension and inferential word learning or word use. Higher scores indicated better ability to comprehend and produce complex, academic language.

The Listening Retell Score (LRS) is derived by adding up the subtotals from four sections:

- Story Grammar (eg. mentioning setting/character/etc.)
- Language Complexity (eg. use of words such as 'because')
- Episode 1 (level of episode complexity dependent on including 2-point problem/attempt/consequence/ending)
- Episode 2 (level of episode complexity dependent on including 2-point problem/attempt/consequence/ending)

Each of these 4 dimensions is present in high quality narratives. The LRS range for the selected benchmark story is 0-47.

The NLM Story Questions [range 0-14] and Vocabulary Questions sections [range 0-9] are supplemental scores. There are 7 Story Questions, each worth a maximum of 2 points. They were designed as a secondary measure of comprehension, particularly for instances where students have not produced a clear or complete narrative retell. There are 3 Vocabulary Questions each worth a maximum of 3 points for clear responses. These questions were designed to measure a student's knowledge of low frequency words and/or to measure a student's ability to infer the meaning of words from context.

Target population: Foundation and Grade 1 Students enrolled in Public and Catholic schools with language and cognitive domain vulnerability rates of ≥ 10% and ICSEA ≤1100 in inner city and regional Victoria.

Treatment conditions: RTI model (intervention period) vs business as usual (control period)

Endpoint: a) CUBED LRS score at the beginning of Grade 2
b) CUBED NLM Story Questions score at the beginning of Grade 2
c) CUBED Vocabulary Questions score at the beginning of Grade 2

Population-level summary measure: Between group (intervention vs. control) adjusted mean difference

Possible Intercurrent Events: As described for the primary outcome (see Section 5.1).

Analysis: The analysis will be similar to the one of the primary outcome (see Section 5.1).

Supplementary Analyses: The analysis will be similar to the one of the primary outcome (see Section 5.1).

5.4 Secondary Outcome: Test of Word Reading Efficiency–Second Edition (TOWRE 2)

Description: TOWRE 2 is a measure of a student's ability to pronounce printed words (Sight Word Efficiency) and phonemically regular non-words (Phonemic Decoding Efficiency) accurately and fluently. The TOWRE 2 provides an efficient means of monitoring the growth of two kinds of word reading skill that are critical in the development of overall reading ability. There are two types of scaled scores: based on student grade, and based on student age (see Stata syntax in APPENDIX D).

- The Sight Word Efficiency (SWE) subtest assesses the number of real words printed in vertical lists that an individual can accurately identify within 45 seconds. A higher score indicates more real words read.
    - Raw score range 0 – 104.
    - Grade Standard scores based on student grade are derived according to the conversion in the table below for Grade 2, range 54-139.
    - Age Standard scores based on student age are derived according to the conversion in the table below

- The Phonemic Decoding Efficiency (PDE) subtest measures the number of pronounceable non-words presented in vertical lists that an individual can accurately decode within 45 seconds. A higher score indicates more pronounceable non-words read.
    - Raw score range: 0 – 66.
    - Grade Standard scores based on student grade are derived according to the conversion in the table below, range 54-143.
    - Age Standard scores based on student age are derived according to the conversion in the table below.

Target population: Foundation and Grade 1 Students enrolled in Public and Catholic schools with language and cognitive domain vulnerability rates of ≥ 10% and ICSEA ≤1100 in inner city and regional Victoria.

Treatment conditions: RTI model (intervention period) vs business as usual (control period)

Endpoint:
a) TOWRE 2 SWE raw score at the beginning of Grade 2
b) TOWRE 2 SWE age standard score at the beginning of Grade 2
c) TOWRE 2 SWE grade standard score at the beginning of Grade 2
d) TOWRE 2 PDE raw score at the beginning of Grade 2
e) TOWRE 2 PDE age standard score at the beginning of Grade 2

f) TOWRE 2 PDE grade standard score at the beginning of Grade 2
g) TOWRE 2 TOTAL raw score (SWE raw+ PDE raw) at the beginning of Grade 2

Population-level summary measure: Between group (intervention vs. control) adjusted mean difference

Possible Intercurrent Events: As described for the primary outcome (see Section 5.1).

Analysis: The analysis will be similar to the one of the primary outcome (see Section 5.1).

Supplementary Analyses: The analysis will be similar to the one of the primary outcome (see Section 5.1).

5.5 Secondary Outcome: Sutherland Phonological Awareness Test – Revised (SPAT - R)

Description: SPAT-R is a measure of a student's phonological and phonemic awareness, which assesses the identification and manipulation of syllables, rhymes and phonemes, and includes tests of non-word reading and spelling. Measures of phonological awareness correlate with concurrent reading and spelling skills and also predict later reading and spelling ability (Neilson, 2003).

In this study 6 subtests of the SPAT-R were administered:

- Blending CVC
- Onset Identification
- Final Phoneme ID
- Segmentation CVC
- Segmentation Blends
- Deletion - Onset

These auditory subtests cover a range of skills measuring student progress in phonological awareness. Each subtest contains 4 items, therefore the total score range for the 6 subtests is 0-24. A higher score indicates greater phonological awareness skills.

Target population: Foundation and Grade 1 Students enrolled in Public and Catholic schools with language and cognitive domain vulnerability rates of ≥ 10% and ICSEA ≤1100 in inner city and regional Victoria.

Treatment conditions: RTI model (intervention period) vs business as usual (control period)

Endpoint: SPAT-R total scores at the beginning of Grade 2

Population-level summary measure: Between group (intervention vs. control) adjusted mean difference

Possible Intercurrent Events: As described for the primary outcome (see Section 5.1).

Analysis: The analysis will be similar to the one of the primary outcome (see Section 5.1).

Supplementary Analyses: The analysis will be similar to the one of the primary outcome (see Section 5.1).

6. REFERENCES

1. Kahan BC, Li F, Copas AJ, Harhay MO. Estimands in cluster-randomized trials: choosing analyses that answer the right question. International Journal of Epidemiology. 2022;52(1):107-18.
2. Gruber JS, Arnold BF, Reygadas F, Hubbard AE, Colford JM, Jr. Estimation of Treatment Efficacy With Complier Average Causal Effects (CACE) in a Randomized Stepped Wedge Trial. American Journal of Epidemiology. 2014;179(9):1134-42.

APPENDIX A: TABLES TEMPLATE

	No of students in each cohort/school and period (No for whom data were available)		
	Period 0	Period 1	Period 2
Cohort 1	XX (XX)	XX (XX)	XX (XX)
…school 1A	XX (XX)	XX (XX)	XX (XX)
…school 1B	XX (XX)	XX (XX)	XX (XX)
…			
Cluster 2	XX (XX)	XX (XX)	XX (XX)
…school 2A	XX (XX)	XX (XX)	XX (XX)
…school 2B	XX (XX)	XX (XX)	XX (XX)
…			

Example Table 1: Students Demographics

	Total		School A		…	
	Control	Intervention	Control	Intervention	Control	Intervention
Number of grade 1 students	XXX	XXX	XXX	XXX	XXX	XXX
Mean Age at assessment (SD)	XX (XX)	XX (XX)	XX (XX)	XX (XX)	XX (XX)	XX (XX)
Gender						
Female	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Male	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
Null/Not Recorded/Other	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)	XX (XX%)
missing	XXX	XXX	XXX	XXX	XXX	XXX

Example Table 2: Description of schools at enrolment

	Total
Number schools	XXX
Location	XX (XX)
North-East region of DE	XX (XX%)
North-West region of DE	XX (XX%)
MACS	XX (XX%)
Language and cognitive domain vulnerability rates	[Median (Range)]
ICSEA	[Median (Range)]
Number of Consent Teachers Per School	[Median (Range)]
Number of Consent students Per School	[Median (Range)]

Example Table 3: Engagement with the intervention

	Total
School/teacher/leaders attendance at the training	XX
School/teacher/leaders attendance at the community of practice sessions	XX
School/teacher/leaders attendance at reflection and planning meetings	XX
N weeks of the year that ISP went to the school	XX

Example Table 4: Primary Outcome

	Control		Intervention	
Number of students	XXXX		XXXX	
RPT	n	Mean(SD)	n	Mean(SD)
mean difference (Intervention - Control)			xx	(95%CI xx, XX)
P-value				0.XXX
Sensitivity analyses:				
1. sensitivity analysis 1				
Mean difference (Intervention - Control)				
P-value				
2. sensitivity analysis 2				
Mean difference (Intervention - Control)				
P-value				
... sensitivity analysis X				
Supplementary analysis:				
Limiting to schools who engaged with the intervention (CACE)				
RPT	n	Mean(SD)	n	Mean(SD)
mean difference (Intervention - Control)			xx	(95%CI xx, XX)
P-value				0.XXX

APPENDIX B: STATA SYNTAX RPT

Step 1 – Remove decimal places from the ages

- Remove decimal places for children aged 6;6-6;11, 7, and 8.
- This cannot be done for children aged 6;0-6;5 because they use a different table for the CELF-5 scaled scores (different table to children aged 6;6-6;11). These will need to be put in manually.

Remove decimal places from numbers

```
gen new_age=int(s_age)
```

Step 2 – Generate scaled/standard scores

Generate RPT standard score

```
recode rpt_raw (0/5=70) (6=71) (7=73) (8=75) (9=76) (10=78) (11=79)
(12=80) (13=81) (14=82) (15=83) (16=84) (17=85) (18=86) (19=88)
(20=89) (21=91) (22=92) (23=94) (24=96) (25=98) (26=100) (27=103)
(28=106) (29=110) (30=117) (31=124) (32=130), generate(rpt_ss)
```

Descriptions RPT

```
recode rpt_ss (70/73=1 "well below average")(74/88=2 "below
average")(89/111=3 "average")(112/126=4 "above average")(127/130=5
"well above average), generate(rpt_ss_des)
```

APPENDIX C: STATA SYNTAX CELF FOLLOWING DIRECTIONS

CELF scaled scores

- Make sure decimal places are removed for ages to just have as 6, 7, 8, 9 etc. as month does not impact scaled score. This will generate new_age

CELF 6;0-6;5

- Scaled scores need manual entering because there are two tables for children aged 6 (one 6;0-6;5 and another for 6;6-6;11). Most children in the study will be 6;6-6;11.

CELF 6;6-6;11

```
recode celf_fd_raw (0=1) (1/2=2) (3=3) (4=4) (5/6=5) (7/8=6)
(9/10=7) (11/12=8) (13/14=9) (15/16=10) (17/18=11) (19=12)
(20/21=13) (22/23=14) (24/25=15) (26=16) (27/28=17) (29/30=18)
(31/33=19) if (new_age==6), gen(celf_ss_6)
```

CELF 7 years

```
recode celf_fd_raw (0/1=1) (2=2) (3/4=3) (5/7=4) (8/9=5) (10/11=6)
(12/13=7) (14/15=8) (16/17=9) (18/19=10) (20=11) (21/22=12)
(23/24=13) (25=14) (26/27=15) (28=16) (29=17) (30/31=18) (32/33=19)
if (new_age==7), gen(celf_ss_7)
```

CELF 8 years

```
recode celf_fd_raw (0/1=1) (2/4=2) (5/7=3) (8/10=4) (11/12=5)
(13/14=6) (15/16=7) (17/18=8) (19/20=9) (21/22=10) (23/24=11)
(25=12) (26/27=13) (28=14) (29=15) (30=16) (31=17) (32=18) (33=19)
if (new_age==8), gen(celf_ss_8)
```

After CELF scaled scores are generated - Combine 6 + 7 + 8 years standard score columns

```
egen celf_ss = rowtotal (celf_ss_6 celf_ss_7 celf_ss_8)
```

Descriptions CELF scaled scores

```
recode celf_ss (1/7=1 "below average")(8/12=2 "average")(13/19=3
"above average"), generate(celf_ss_des)
```

APPENDIX D: STATA CODING FOR TOWRE 2 STANDARD SCORES DERIVATION

---- Grade standard score

TOWRE **SWE Grade** standard score

```
recode towre_swe_raw (0=54) (1=55) (2/3=56) (4=57) (5/6=58) (7/9=59)
(10/12=60) (13/15=61) (16/19=62) (20=66) (21=69) (22=70) (23=71)
(24/25=72) (26/27=74) (28=75) (29=76) (30=79) (31/35=80) (36=81)
(37/38=82) (39=83) (40/41=84) (42=85) (43=86) (44=88) (45=89)
(46=90) (47=91) (48=92) (49=93) (50=96) (51=97) (52=98) (53/54=100)
(55=102) (56=104) (57/58=105) (59=106) (60=108) (61=109) (62=112)
(63/64=114) (65=116) (66=117) (67=118) (68=121) (69=123) (70=125)
(71=126) (72=131) (73=138) (74/108=139), generate(towre_swe_grst)
```

Descriptions TOWRE **SWE Grade** standard score

```
recode towre_swe_grst (54/69=1 "very poor")(70/79=2 "poor")(80/89=3
"below average")(90/110=4 "average")(111/120=5 "above
average")(121/130=6 "superior")(131/139=7 "very superior"),
generate(towre_swe_grst_des) test
```

TOWRE **PDE Grade** standard score

```
recode towre_pde_raw (0=54) (1=60) (2=63) (3=67) (4=70) (5=72)
(6=77) (7/8=80) (9=82) (10=84) (11=86) (12=88) (13=89) (14=91)
(15=93) (16=94) (17=95) (18=96) (19=98) (20=99) (21=100) (22=101)
(23=103) (24=105) (25=106) (26=107) (27=108) (28=109) (29=110)
(30/31=113) (32=114) (33=116) (34=117) (35=118) (36=119) (37=120)
(38/39=122) (40=125) (41=126) (42=127) (43=129) (44=133) (45=134)
(46=135) (47=137) (48=139) (49=142) (50/66=143), generate
(towre_pde_grst)
```

Descriptions TOWRE **PDE Grade** standard score

```
recode towre_pde_grst (54/69=1 "very poor")(70/79=2 "poor")(80/89=3
"below average")(90/110=4 "average")(111/120=5 "above
average")(121/130=6 "superior")(131/139=7 "very superior"),
generate(towre_pde_grst_des) test
```

---- Age standard score SWE

6 years

```
recode towre_swe_raw (0=54) (1=64) (2=68) (3=70) (4=72) (5=73)
(6=74) (7=75) (8=76) (9=78) (10=79) (11/12=80) (13=81) (14=82)
(15/16=83) (17=84) (18=85) (19=86) (20=87) (21/24=88) (25/27=89)
(28/30=90) (31=92) (32=93) (33/34=94) (35/36=95) (37/38=96) (39=101)
(40=102) (41/42=104) (43=105) (44/45=106) (46=107) (47/48=108)
(49=109) (50=110) (51/52=111) (53/55=112) (56=113) (57/58=114)
(59=118) (60=120) (61/62=121) (63/64=122) (65=123) (66=124) (67=128)
(68=130) (69=132) (70=134) (71=138) (72/108=139) if (new_age==6),
generate(towre_swe_agest_6)
```

7 years

```
recode towre_swe_raw (0=54) (1=56) (2=58) (3=60) (4=61) (5/6=62)
(7=63) (8/9=64) (10/11=65) (12/13=66) (14=73) (15=74) (16=75)
```

```
(17=76) (18=77) (19=78) (20/21=79) (22/24=80) (25/27=81) (28=82)
(29/32=84) (33/34=85) (35=86) (36=87) (37=89) (38=90) (39=92)
(40/42=94) (43=96) (44/46=97) (47=98) (48=99) (49=100) (50=101)
(51=103) (52=104) (53=105) (54=107) (55=109) (56=111) (57=112)
(58=114) (59=115) (60/61=116) (62=118) (63/64=121) (65=122) (66=123)
(67=124) (68=126) (69=127) (70=128) (71=130) (72=131) (73=134)
(74/108=135) if (new_age==7), generate(towre_swe_agest_7)
```

8 years

```
recode towre_swe_raw (0/1=54) (2=55) (3/4=56) (5/6=57) (7/9=58)
(10/12=59) (13/16=60) (17/20=61) (21/25=62) (26=69) (27=71) (28=72)
(29=73) (30=75) (31=76) (32/33=78) (34/36=79) (37/39=80) (40/41=81)
(42/43=83) (44=84) (45=86) (46=87) (47/48=89) (49=90) (50=92)
(51=93) (52=96) (53=97) (54/55=98) (56=99) (57=100) (58=101)
(59=102) (60=104) (61=105) (62=107) (63=109) (64=110) (65=111)
(66=112) (67=113) (68=115) (69=117)(70=120) (71=121) (72=124)
(73=125) (74=128) (75=129) (76=130) (77=132) (78=133) (79=134)
(80=135) (81=137) (82=138) (83=140) (84=144) (85/108=145) if
(new_age==8), generate(towre_swe_agest_8)
```

Combine 6 + 7 + 8 years standard score columns

```
egen towre_swe_agst = rowtotal (towre_swe_agest_6 towre_swe_agest_7
towre_swe_agest_8)
```

Descriptions TOWRE SWE age standard score

```
recode towre_swe_agest (54/69=1 "very poor")(70/79=2 "poor")(80/89=3
"below average")(90/110=4 "average")(111/120=5 "above
average")(121/130=6 "superior")(131/139=7 "very superior"),
generate(towre_swe_agest_des)
```

---- Age standard score PDE

6 years

```
recode towre_pde_raw (0=54) (1=72) (2=74) (3=78) (4=84) (5=85)
(6=88) (7/8=92) (9/10=94) (11=95) (12=96) (13=99) (14=100) (15=102)
(16=104) (17=105) (18/20=106) (21=107) (22=108) (23=109) (24=110)
(25=111) (26=113) (27=114) (28/29=115) (30=116) (31=118) (32/33=119)
(34=120) (35=122) (36/37=124) (38=125) (39/40=126) (41=127) (42=128)
(43=130) (44=131) (45=133) (46=137) (47/66=138) if (new_age==6),
generate(towre_pde_agest_6)
```

7 years

```
recode towre_pde_raw (0=54) (1=66) (2=70) (3=73) (4=74) (5=77)
(6=80) (7=82) (8=83) (9=85) (10=88) (11=90) (12=91) (13=92) (14=93)
(15=96) (16/17=97) (18=98) (19/20=100) (21=102) (22=103) (23=105)
(24=107) (25=109) (26=110) (27=111) (28/29=112) (30=114) (31=116)
(32=117) (33=119) (34/35=120) (36/37=122) (38/39=123) (40=125)
(41/42=126) (43=127) (44=130) (45=131) (46=133) (47=134) (48=138)
(49=141) (50/66=142) if (new_age==7), generate(towre_pde_agest_7)
```

8 years

```
recode towre_pde_raw (0=54) (1=67) (2=71) (3=72) (4=73) (5=75)
(6=78) (7/8=81) (9=83) (10=84) (11=85) (12=87) (13=88) (14=91)
(15=92) (16/17=93) (18=95) (19/20=97) (21=98) (22=99) (23=100)
(24=101) (25/26=102) (27=104) (28=105) (29=106) (30=107) (31/32=109)
(33/34=112) (35/36=114) (37/38=116) (39=118) (40/41=121) (42=124)
(43=128) (44=129) (45=133) (46=135) (47=136) (48=138) (49=141)
(50/66=142) if (new_age==8), generate(towre_pde_agest_8)
```

Combine 6 + 7 + 8 years standard score columns

```
egen towre_pde_agst = rowtotal (towre_pde_agest_6 towre_pde_agest_7
towre_pde_agest_8)
```

Descriptions TOWRE PDE age standard score

```
recode towre_pde_agest (54/69=1 "very poor")(70/79=2 "poor")(80/89=3
"below average")(90/110=4 "average")(111/120=5 "above
average")(121/130=6 "superior")(131/139=7 "very superior"),
generate(towre_pde_agst_des)
```