

Cambodia

SMILE PROGRAMME

EVALUATION 2022–2026

Composite Report | June 2026



Publication Information

Published by:

© 2026 VVOB – *education for development*

Julien Dillensplein 1 bus 2A, 1060 Brussels, Belgium

Telephone: +32 2 209 07 99; Website: www.vvob.org

Some rights reserved.

This report was prepared by BN Consult, with Jeffery H. Marshall engaged as the lead independent researcher, M. Alejandra Sorto as the maths education expert, and commissioned by VVOB in Cambodia under the SMILE Programme.

The findings, interpretations and conclusions expressed in this report are those of the external evaluation team and do not necessarily reflect the views of VVOB, the Ministry of Education, Youth and Sport, or the programme's funding partners.

Rights and Permissions

Responsible Editor:

Tom Vandenbosch

© 2026 VVOB – *education for development*



This work is available under the Creative Commons Attribution-Non-Commercial Share-alike 4.0 International license (CC BY-NC-SA 4.0). Under the Creative Commons Attribution license, you are free to copy, distribute, transmit, and adapt this work, only for non-commercial purposes, under the following conditions:

- **Attribution**—Please cite the work as follows: Marshall, J.H. (2026). SMILE Programme Evaluation 2022–2026: A Longitudinal Quasi-Experimental Study. VVOB – *education for development*.
- **License**: Creative Commons Attribution-NonCommercial Share-Alike 4.0 International license (CC BY-NC-SA 4.0)
- **Translations**—If you create a translation of this work, please add the following disclaimer along with the attribution: This translation was not created by VVOB and should not be considered an official VVOB translation. VVOB shall not be liable for any content or error in this translation.
- **Adaptations**—If you create an adaptation of this work, please add the following disclaimer along with the attribution: This is an adaptation of an original work by VVOB. Views and opinions expressed in the adaptation are the sole responsibility of the author or authors of the adaptation and are not endorsed by VVOB.
- **Distribution**—Licensees may distribute derivative works only under a license identical to the license that governs the original work.
- **Third-party content**—VVOB does not necessarily own each component of the content contained within the work. VVOB therefore does not warrant that the use of any third party-owned individual component or part contained in the work will not infringe on the rights of those third parties. The risk of claims resulting from such infringement rests solely with you. If you wish to re-use a component of the work, it is your responsibility to determine whether permission is needed for that re-use and to obtain permission from the copyright owner. Examples of components can include, but are not limited to, tables, figures, or images.

All queries on rights and licenses should be addressed to VVOB, Julien Dillensplein 1 bus 2A, 1060 Brussels, Belgium

- Telephone: +32 2 209 07 99; E-mail: info@vvob.org; Website: www.vvob.org

Disclaimer

This publication contains references to other publications and websites. VVOB cannot be held responsible for current or future content of these publications and websites, nor for the content of external publications and websites that refer to this publication of VVOB.

This publication was produced with the financial support of the Government of Belgium.

Its contents are the sole responsibility of VVOB and do not necessarily reflect the views of these actors.

The Government of Belgium cannot be held responsible for the content of this publication.

Executive Summary

SMILE Programme and study background

Since 2022 VVOB (*education for development*) has worked with the Cambodia Ministry of Education, Youth and Sports (MoEYS) to implement a five-year programme called “Strengthening early grade Maths through Inclusive Level-appropriate Education” (SMILE) in three provinces (Kampot, Kep and Kampong Thom). The SMILE programme aims to improve early grade math (EGM) teaching and learning by strengthening teacher training on EGM pedagogical content knowledge, level appropriate gender-responsive teaching, and classroom management.

The SMILE programme implementation during 2022–2026 has been accompanied by a rigorous, longitudinal data collection and analysis to provide evidence on programme impact for key student and teacher outcomes. The design relies on a matched set of 100 schools (50:50) drawn from Kampong Thom (treatment) and Pursat (control) provinces. The student and teacher data has been extensively analysed throughout the study period, with detailed reports covering the teacher instrument piloting phase and the baseline and endline data collections for students and teachers in grades 1, 2 and 3 (a total of seven reports). Each report includes a sizeable amount of descriptive statistics and multivariate analysis results, with very detailed methodological background summaries.

Analytical framework

The overall study design and data analysis address three research questions that focus on SMILE programme impact on the teacher’s (self-reported) efficacy in early grade mathematics teaching, their pedagogical content knowledge (PCK), and their early grade student learners’ mathematics achievement progress. The matched 50:50 school design was created by first randomly selecting 50 schools from five districts in Kampong Thom (SMILE programme province). These schools were then matched with 50 similar schools from five districts in Pursat (control province) using “propensity score” matching as well as other school characteristics.

A cohort of nearly 1,000 students was tested from the beginning of grade 1 (November 2023) through the end of grade 3 (November 2024) using the MoEYS-developed Early Grade Mathematics Assessment (EGMA) tool. Data from SMILE programme teachers (Kampong Thom) and controls (Pursat) in grades 1 (2023), 2 (2024) and 3 (2025) were collected at the beginning (pre-training) and end of each school year using two sets of instruments: 1) self-reported teaching activities, capacity and attitudes; and 2) a test of pedagogical content knowledge (PCK). The research team was responsible for the application of the teacher instruments, while the student assessment was conducted by district and provincial MoEYS staff with experience in EGMA, with refresher training and supervision provided by EQID.

The teacher and student data were subjected to a very detailed, comprehensive analysis that included a series of diagnostic activities to assess construct stability and overall instrument efficiency, cohort attrition and balance, and the robustness of the main findings for teachers and students. The diagnostic results are extensive and are detailed in the individual reports. The teacher instruments demonstrated a high degree of construct stability and reliability across the various data collection moments, the teacher and student samples had similar rates of attrition with limited evidence of divergent attrition patterns between treatment and control samples (e.g. “selective attrition”), and the main findings from the data analysis were generally robust to various diagnostic checks.

The key report outcome includes the treatment effects estimations that attempt to isolate the impact of SMILE programme participation on the teacher and student outcomes of interest. It is important to be clear that the absence of random assignment means that the treatment effects can only be estimated using statistical modelling applied to a matched design, and these results cannot be assumed to be strictly causal. Nevertheless, with three separate cohorts of teachers (grades 1, 2 and 3), and a large cohort of students with a panel of four measures of mathematics knowledge, the data provide numerous possibilities for comparing treatment and control study subjects. The various methods include difference-in-difference (DID), individual matching (for students) and combined (“pooled”) multivariate regression modelling incorporating random and fixed effects.

Main findings: Teachers

The results from the teacher data analysis are mixed, and highlight the importance of a robust statistical methodology and a detailed review of the teacher results for the overall constructs as well as specific content areas. For the teacher's self-reported efficacy in early grade mathematics teaching the results show that the endline averages are generally higher than the (pre-treatment) baseline averages, and in most cases the treatment and control teacher samples had very similar baseline averages. There is limited evidence that treatment teachers experienced larger improvements between baseline and endline on the various teacher survey question blocks. The main exception is in survey section 2 (Maths content-pedagogy self-efficacy) where in all three grades the treatment teachers have higher or equal endline scores despite starting with lower (or equal) scores at baseline. The larger improvements in section 2 for treatment teachers—combined with very similar averages on the other teacher survey sections—result in marginally higher rates of improvement for treatment teachers on the overall teacher survey in all three grades, although again with the caveat that endline scores are virtually identical between control and treated in all comparisons.

The difference-in-difference modelling (treatment effects) goes a step further than the descriptive statistic comparison of baseline and endline results. The headline finding is that treatment teachers have significantly larger gains between baseline and endline for the overall teacher survey instrument, as well as the Section 2 (Maths content-pedagogy) and (to a lesser degree) Section 3 content (Classroom management). The effect sizes are fairly large, with an overall gain score advantage of 0.30–0.44 standard

deviations, and an even larger effect for the Section 2 content (0.50–0.60 SD).

The evidence of a SMILE programme effect for the overall survey and especially section 2 is a positive result that is generally consistent across all three data collection and analysis periods. However, there are some caveats to be noted. First, there is some evidence of non-random attrition patterns that potentially affect the comparisons of these teacher survey indicators at endline. Also, the significant programme impact is concentrated in the “Maths content and pedagogy” section—especially “Training application”—and reflects a higher rate of improvement for treatment teachers from a lower baseline average in two of the three teacher cohorts. The fact that treatment teachers reported more improvement in their confidence in geometry, measurement and especially the application of training content is an important finding, and is consistent with programme intent. And this improvement in questionnaire section 2 content is large enough to generate a fairly large treatment effect for the overall teacher questionnaire average. Nevertheless, the bottom line is that the treatment effect for Training Application difference is the only robust result among the various significant differences, and when removing this content area there is much less evidence of a SMILE programme impact on the overall teacher self-efficacy instrument.

There is no evidence of VVOB programme impact on teacher pedagogical content knowledge (PCK). In all three grades the baseline and endline scores are very similar, and the same is true for the net change between baseline and endline (i.e. gain score). The actual changes between baseline and endline are difficult to interpret due to modifications to the test content between baseline and endline, although this is mainly the case for grade 1.

Main findings: Students

The use of a single student cohort with as many as four observations per student means there are more analysis options available. Descriptive statistics comparisons show that treatment sample scores are significantly higher than control sample scores in almost all of the comparisons in later stages of data collection when most students were in grades 2 (2024) and 3 (2025). For example, treatment and control groups both answered about six percent of Addition 1 content questions correctly in the grade 1 pre-treatment baseline, but by the 2025 endline the treatment student advantage was nearly 10 percent (65% versus 56%). In standard deviations (“z-score”)

the treatment student advantage on the overall EGMA test is 0.30 standard deviations at the end of 2024, and 0.32 SD at the end of 2025.

The VVOB student advantage is generally larger in more advanced mathematics topics, including the Number Pattern, Addition 2 and Subtraction 2 content areas that were added to the 2024 and 2025 endline EGMAs. By contrast, on easier content there is some evidence of programme impact after the initial baseline test, but for Number Identification and Number Discrimination the overall averages were above 90 percent correct by the 2025 endline, so there is not much room for a programme effect.

The treatment effects results are generally consistent with the more basic descriptive statistics summaries, and show large (and significant) VVOB programme advantages in Addition 1, Subtraction 1 and the overall score. The effect size for the overall EGMA scores are somewhat smaller, especially in the regression-based estimations: for example, the overall standardised difference based on percentage correct is 0.29 standard deviations in the basic DID, and 0.19 SD when using multivariate modelling (see “Model 1” and “Model 2”).

The main findings for the difference-in-difference approach are generally robust to the different diagnostic corrections for multiple hypothesis testing and student attrition, and the results based on fixed effects modelling also show relatively large (and significant) treatment effects in Addition 1 and Subtraction 1 and smaller, but significant, effects for overall EGMA test measures.

Summary discussion

The cumulative results from three years of data collection and analysis support two main findings. First, there is substantial evidence of a significant VVOB programme impact on student achievement in Mathematics. And second, there is limited evidence of a programme impact on the teacher efficacy survey content, and no evidence of impact on teacher pedagogical content knowledge (PCK).

The evidence of a VVOB programme impact on student achievement is encouraging, but the source of this apparent treatment effect is not clear. Neither of the two main teacher indicators (self efficacy survey and PCK) are consistently significant in the modelling of student achievement, so even the modest programme impact on the teacher survey content cannot be directly linked with student outcomes. In other words, the underlying mechanisms that have led to the apparent programme impact on student achievement are not identified through the data collection. That work remains for additional evaluations to address, such as the classroom observation study or qualitative data that unpacks programme components to understand more about the implementation challenges and successes.

These findings are consistent with recent research in Cambodia linking the Early Grade Learning (EGL) package—which the VVOB mathematics intervention is part of—with improvements in student learning. Most of this evidence is on EGMA rather than EGMA, and the comparability of the designs and reporting is somewhat limited. But the important point is simply that the VVOB evidence is clearly contributing to a growing literature on

The main takeaway from the student assessment treatment effects analysis—which is the most important aspect of the VVOB programme evaluation—is that there is clear evidence of programme impact. The treatment and control students began the treatment period with very similar levels of achievement, but by the end of the testing period the VVOB students were clearly outperforming their control school counterparts. Furthermore, the treatment and control samples have similar attrition dynamics, which enhances the comparability of the samples across all the data collection periods. The effect sizes are relatively large in individual test content areas (upwards of 0.50 SD or more), but the overall EGMA test performance metrics provide the best measures for gauging programme impact. These results are somewhat variable, but the overall programme impact is consistently in the 0.10–0.20 standard deviation range based on the most rigorous treatment effects estimators.

EGL dynamics in Cambodia, which is critical for continuing the work of improving student learning outcomes through more prepared and supported teachers.

The main findings are also in line with international evidence, although these comparisons are even more problematic given differences in interventions, research designs and contexts. The main takeaways from the international literature are clearly relevant for the Cambodian teacher training context. Teacher professional development interventions have been linked with student achievement gains, but they tend to be fairly modest and are larger when the intervention is content-focused, includes active learning, provides ongoing coaching, and is embedded in the teacher’s daily work. The VVOB programme has some of these features, but there are clear challenges in delivering high quality training and support through the Training of Trainers (ToT) approach that relies heavily on government counterparts.

Finally, the cost effectiveness analysis (CEA) suggests that the VVOB programme compares somewhat favourably with other interventions (international evidence base), but the results are quite sensitive to model specification and calculation inputs. The costs to improve student achievement by 0.10 SD are between \$5.15 and \$8.79, which is similar to the CEA results from a government-led TaRL programme in Zambia (\$6.38 per 0.10 SD*). Another set of CEA parameters finds that the VVOB programme produces between 1.1 and 1.9 standard deviations of learning for each \$100 of programme spending. JPAL has shown that quite a few programmes globally are in the 1.0–2.0 SD range, but with some very large exceptions (larger and smaller).

* De Barros, A. and T. Lubozha (2026). *Targeting foundational skills at scale: Skill specificity and transfer*. CES IFO Working Papers 12542.

TABLE OF CONTENTS



3. Main findings: Teachers 17

3.1 Construct stability	18
3.2 Attrition and balance comparisons	20
3.3 Treatment effects and diagnostics	22
3.4 Predictors of teacher outcomes	24
3.5 Gender dimensions	25
3.6 Formative feedback	26

4. Main findings: Students 28

4.1 Attrition and balance	29
4.2 Treatment effects and diagnostics	31
4.3 Predictors of student test scores	35
4.4 Gender dimensions	35



1. Introduction 08

2. Analytical framework 10

2.1 The SMILE programme	11
2.2 Research questions	12
2.3 Research design and sampling	12
2.4 Instrumentation	13
2.5 Data collection	14
2.6 Data analysis methods	15



5. VVOB programme effectiveness discussion 37

5.1 Recent evidence from Cambodia	38
5.2 International evidence	39
5.3 Cost-effectiveness analysis (CEA)	40

6. Main findings and recommendations 43

Report bibliography 47

Data Annex A 48



01

INTRODUCTION

Introduction

Since 2022 VVOB (Education for Development) has worked with the Cambodia Ministry of Education, Youth and Sport (MoEYS) to implement a five-year programme called “Strengthening early grade Maths through Inclusive Level-appropriate Education” (SMILE) in three provinces (Kampot, Kep and Kampong Thom). The SMILE programme aims to improve early grade math (EGM) teaching and learning by strengthening teacher training on EGM pedagogical content knowledge, level-appropriate, gender-responsive teaching, and classroom management. The intervention includes a combination of training, mentoring and peer learning. The ultimate programme goal is to ensure that boys and girls in early primary grades (1–3) gain relevant and effective knowledge in mathematics that prepares them for future learning.

The SMILE programme implementation during 2022–2026 has been accompanied by a rigorous, longitudinal data collection and analysis to provide evidence on programme impact for key student and teacher outcomes. The design relies on a matched set of 100 schools (50:50) drawn from Kampong Thom (treatment) and Pursat (control) provinces. A cohort of nearly 1,000 students was tested from the beginning of grade 1 (January 2023) through the end of grade 3 (August 2025) using the MoEYS-developed Early Grade Mathematics Assessment (EGMA) tool. SMILE programme teachers (Kampong Thom) and controls (Pursat) in grades 1 (2023), 2 (2024) and 3 (2025) completed instruments related to teaching activities, capacity and attitudes, and also took a test of their pedagogical content knowledge, before each school year and then at the end of the school year.

The student and teacher data have been extensively analysed throughout the study period, with detailed reports covering the teacher instrument piloting phase and the baseline and endline data collections for students and teachers in grades 1, 2 and 3 (a total of seven reports, see the Report bibliography). Each report includes a sizeable amount of descriptive statistics and multivariate analysis results, with very detailed methodological background summaries.

This summary report provides an overview of the evidence that has been generated during the four years of analysis. The composite summary is necessary to provide a general overview of the results in an accessible format that focuses on the main findings. This includes brief reviews of the SMILE programme, study and sample design, methodological and data collection details, and data analysis results. The focus is on the main study findings and interpretation. Interested readers can consult the reports from each data collection moment for more detailed summaries.

Section 2 reviews the SMILE programme, study design and research questions, sampling plan and data analysis methods. Section 3 reviews the main findings from the two teacher instruments (teacher survey and pedagogical content knowledge test), while Section 4 reviews the main findings from the student EGMA testing cohort analysis. Section 5 considers how the SMILE programme effectiveness evidence compares with other relevant studies and programmes, including cost-effectiveness dimensions. Section 6 concludes with the main findings and recommendations.



02

**ANALYTICAL
FRAMEWORK**

Analytical Framework

2.1/ The SMILE Programme

National assessment evidence from Cambodia has consistently shown that primary students are performing far below expected levels in both language and (especially) mathematics (EQID, various years). For example, the 2020 national assessment carried out by the Education Quality and Inspection Department (EQID) found that grade 3 students were only able to correctly answer 40 per cent of the test questions. This assessment also found that boys outperformed girls in mathematics.¹

The SMILE programme intervention is designed to address two general systemic deficiencies that affect teaching and learning outcomes in mathematics. **First, low levels of pedagogical content knowledge (PCK) among teachers are a likely contributor to low student learning outcomes.** PCK goes beyond the knowledge of mathematical content and refers to a variety of competencies of teachers to teach a specific subject well: selecting the right examples, asking good questions, understanding and remediating mistakes learners make, integrating learning resources effectively in lessons, etc.² **The second area of concern is the pervasive gender stereotypes and gender blindness in maths classes.** Girls are asked fewer or less challenging questions, they are assigned less responsible roles in group activities, and expectations for girls are lower in general.³

The overarching goal of the VVOB SMILE programme is to ensure that newly qualified and in-service early grade (primary grades 1–3) teachers provide effective, level-appropriate and gender-responsive early grade maths teaching, thereby improving all learners' math skills and overall wellbeing.

The approach works through close partnerships with the MoEYS and Provincial Teacher Training Centers (PTTC) to strengthen the professional development of teachers and school leaders on early grade mathematics (EGM) in primary education. Programme teachers are expected to acquire relevant knowledge, attitudes and skills to

help them regularly and informally assess learners' needs and adapt their teaching accordingly. This capacity development process is built around 3 pillars:

- Strengthening the competencies of teacher trainers and ministry staff to effectively support school-based mentors, primary school leaders and (student) teachers in teaching EGM in a level-appropriate and gender-responsive way.
- Strengthening pedagogical content knowledge and classroom management skills of teachers through a continuous professional development (CPD) trajectory that is complemented by remote learning and mentoring, allowing teachers to put their newly strengthened skills to practice.
- Establishment of collaboration, dialogue and adaptive learning between MoEYS and (a) research institution(s) to develop an evidence-informed scaling strategy for pre-service and in-service teacher training on level-appropriate and gender-responsive quality EGM teaching.

SMILE programme participant teachers are supported through two main mechanisms. **First, they receive a grade-specific two-week training prior to the school year.** These trainings began in 2022 with grade 1 teachers, and continued with grade 2 (2023) and grade 3 (2024) teachers. **The second mechanism is through school and district-based mentors that work closely with in-service teachers to enhance their skills.**

As part of its revised teacher support system, the MoEYS appoints school- and district-based mentors – usually experienced teachers – to support school leaders in the provision of school-based continuing professional development (CPD) for early grade teachers. The SMILE programme is supporting these efforts by building the capacity of school leaders and mentors to support teachers to participate in CPD for EGM.

¹ EQID 2020 grade 3 national assessment.

² SMILE programme website summary: <https://www.vvob.org/our-work/our-projects/smile> (accessed April 15, 2026).

³ Ibid.

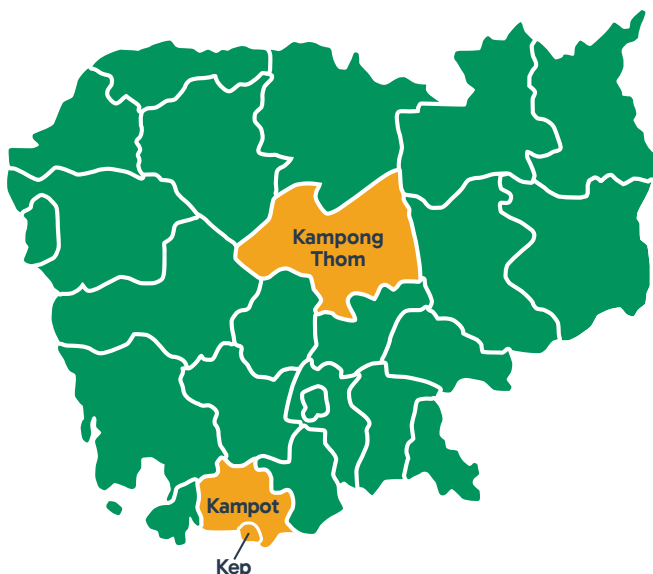
2.2/ Research questions

The quasi-experimental SMILE programme evaluation study aims to measure the impact of the level-appropriate, gender-responsive maths professional development package on in-service teachers (INSET and their Grade 1 to 3 learners). The study addresses three research questions:

- First, does SMILE programme participation impact teacher efficacy regarding effective, level-appropriate and gender-responsive early grade math teaching?
- Second, does SMILE programme participation impact teacher pedagogical content knowledge in mathematics?
- And third, is there evidence that SMILE programme participation has an impact on core mathematic skills of early grade primary school learners?

2.3/ Research design and sampling

The SMILE programme was implemented in all primary schools in three provinces (Kampot, Kampong Thom and Kep). This approach facilitates implementation since the intervention components (trainings carried out by master trainers who are supported by a Core Expert Group, plus school visits) can be carried out by provincial and district officers in all schools (and clusters). The downside to this implementation approach is that it is not possible to randomly assign schools to treatment or control categories within each province (or district) as their treatment status is instead determined solely by geographic location.



The challenge was therefore to find a relevant set of schools to compare with schools that are participating in the SMILE programme. This was done using a matched design where 50 schools were randomly chosen from five districts in one of the SMILE provinces (Kampong Thom), and then matched with 50 schools from a similar, nearby province (Pursat) where the programme is not being implemented. The samples were obtained from five districts in each province to reduce data collection costs while maintaining broad coverage across the eight districts in Kampong Thom, and six districts in Pursat.

The matching and sampling details are provided in a separate report from September 2022.⁴ In the first step the probability of programme participation (called a “propensity score”) was obtained for each school in Kampong Thom and Pursat using a multivariate regression estimation with basic school characteristics (enrolment, infrastructure, etc.) as predictors. Five districts from Pursat were then matched with the five districts from Kampong Thom where the 50 treatment schools were randomly selected based on the average propensity score for the schools in each district. **The 50 matched schools were then chosen on the basis of the propensity score block as well as other school characteristics like the number of grade 1 classes, whether or not the school is a cluster or satellite, urban-rural location, etc.**

⁴. BN Consult (2022). “VVOB-SMILE Programme Evaluation Sampling Note.” Phnom Penh.

The study design requirements (based on the VVOB contract with the data collection firm) called for samples of 200 teachers to be surveyed in each cohort (grades 1–3). Most of the schools from the 50:50 matched sample have only one class in primary grades, so it was not possible to rely exclusively on the 50:50 matched sample to obtain the necessary 200 teachers. **This required “supplementary” schools that were chosen in each province to provide additional teachers to reach the 200 teacher target.** These schools were selected first in the Kampong Thom treatment province by identifying additional schools within the same commune as the original 50 sampled schools. And then appropriate matches were found in the same communes where the 50 matched Pursat schools had been sampled. The supplemental schools are only used to increase the teacher sample size, they are not relevant for the student analysis. In each data collection moment there are 25–30 supplemental schools in each province (Kampong Thom, Pursat).

The teacher sample was determined entirely by the 50:50 matched design, plus the matched supplemental schools. In the first year of the study all grade 1 teachers in the 50 treatment schools from Kampong Thom (plus supplemental schools) were identified during the SMILE teacher training that took place before the 2023 school

year. The grade 1 teachers in the 50 matched schools (plus supplemental schools) from the control province were identified at the beginning of the school year. All selected grade 1 teachers were asked to complete the two teacher data collection instruments (explained in section 2.4). At the end of the grade 1 school year these same grade 1 teachers—or their replacements (replacements are discussed in section 2.5)—were then given the same instruments to complete. This process for teacher selection was then repeated for grade 2 (2024) and grade 3 (2025).

The student sampling was different since it relied on a single cohort of grade 1 students that were chosen at the beginning of the 2023 school year, and this cohort was then followed in 2024 and 2025. **In each school 10 grade 1 students (5 boys, 5 girls) were randomly selected to complete the EGMA assessment, which was conducted as a one-on-one interview by data collection staff.** The student EGMA assessment was only completed in the 50:50 matched schools, not the supplemental schools. In the year 1 baseline (i.e. beginning of grade 1) a total of 987 students (494 from control schools, 493 from treatment) completed the EGMA test. These same students were then followed throughout the course of the longitudinal study.

2.4/ Instrumentation

The study relies on three instruments for measuring SMILE programme impact. The teacher survey was developed by the analysis team in close coordination with VVOB, and includes aspects that are aligned with the SMILE treatment and have been shown to be associated with student performance and classroom practices. The teacher survey development began with a literature review of relevant areas of research on teacher effectiveness⁵. An initial instrument was then developed with eight sections, including: 1) teacher background; 2) self-confidence/efficacy in mathematical content and pedagogy (numbers, measurement, geometry, applying training knowledge); 3) classroom management strategies (positive discipline, differentiation, flexible grouping); 4) knowledge, attitudes and practices (strength of beliefs in capabilities, attributes, mathematics anxiety); 5) awareness of mathematical dispositions; 6) mathematics productive disposition scale (MPDS) (affect, belief, identity, mathematical integrity, risk-

taking, goals, motivation, self-efficacy); 7) gender-responsive pedagogy and 8) formative programme feedback (for Kampong Thom teachers only).

The teacher survey sections include blocks of questions with ordered responses indicating degree of agreement or frequency of usage. The questions were drawn primarily from existing teacher survey instruments that are described in more detail in the literature review document.

The teacher **pedagogical content knowledge (PCK) test** was also developed specifically for the SMILE programme evaluation by the analysis team. The PCK questions are designed to measure specific strands of teacher knowledge that are relevant for teaching, although content knowledge is embedded in the questions. The PCK test was divided into four sections: 1) numbers and operations; 2) fractions and number patterns; 3) geometry and measurement; and 4) problem solving.

⁵ BN Consult (2022). “Teacher questionnaire literature review.” Phnom Penh.

The teacher instruments were piloted in 20 primary schools in Kandal province in November 2022. Kandal was chosen due to its proximity to Kampong Thom and Pursat provinces, and to avoid contamination of the Kampong Thom and Pursat teacher samples. 79 teachers completed the instruments.

The teacher instrument pilot data analysis is summarised in a separate report⁶. The data analysis methods are discussed in more detail in section 2.6 below. The initial versions of the instruments were quite long, so a key objective of the pilot data analysis was to identify questions that could be removed ahead of application in the full grade 1 teacher sample. The teacher survey instrument had 145 questions in the original pilot version, which was reduced to 128 questions after the pilot analysis. The PCK test began with 57 questions and was reduced to 34 questions.

For measuring student mathematics knowledge the Cambodian version of the **Early Grade Mathematics**

Assessment (EGMA) was incorporated in all four testing moments. The student cohort was tested at the beginning of grade 1 in January–February 2023 (before SMILE programme implementation), and then again at the end of grade 1 (November 2023) using the same assessment instrument. The cohort was then tested at the end of grade 2 in December 2024 (or in grade 1 for students who were repeating), and finally at the end of grade 3 in August 2025 (or in grade 1 or 2 for grade repeaters). The EGMA instrument cannot be modified, and the student assessment was carried out by a different team of experienced enumerators that were trained and supported by the Education Quality and Inspection Department (EQID).

The grade 1 EGMA version covered number identification, number discrimination, addition and subtraction. The grade 2 and 3 EGMA version included the same content as the grade 1 version, but with additional sections for number patterns and more advanced addition and subtraction content (“Addition 2”, “Subtraction 2”).

2.5/ Data collection

The data collection details are provided in the reports that were produced for each data collection moment (seven in total including the pilot study). The consulting firm was responsible for the application of the teacher instruments. The student assessment was conducted by district and provincial MoEYS staff with experience in EGMA, with refresher training and supervision provided by EQID. The teacher survey was completed using tablets, while the PCK test was applied in a pencil and paper format. The data collection firm has a cadre of experienced enumerators and supervisors, many of whom participated in multiple rounds of the data collection (or all rounds). The survey firm employed a range of quality control measures for ensuring data quality, and the data were transferred over to Stata and SPSS for data analysis.



There were very few problems encountered during the data collection. One problem in the initial data collection stages was the length of the teacher instruments, which some teachers (especially older teachers) reported were too long. However, this was less of an issue in later stages as the teacher instruments (especially the survey) were further streamlined after the grade 1 baseline and endline data collections.

“The student assessment was conducted by district and provincial MoEYS staff with experience in EGMA, with refresher training and supervision provided by EQID.”

⁶ BN Consult (2023). “Pilot and validation report.” Phnom Penh.

2.6/ Data analysis methods

The teacher and student data were subjected to a very detailed, comprehensive analysis that focused on five main sets of activities. **First, in all survey moments the teacher survey and PCK test data were analysed using factor and reliability analysis to assess construct stability and improve overall instrument efficiency.** The teacher survey includes a large number of constructs and sub-constructs (e.g. the general construct of “mathematical content and pedagogy self-efficacy” and then “numbers” as a sub-category). The analysis was used to continuously reassess the questionnaire content to confirm that the different question blocks were effectively loading onto the same constructs, with sufficient overall factor power (e.g. eigenvalues greater than 1.0) and, when possible, to further streamline the teacher survey by removing items (or entire blocks) with smaller factor loadings and lower levels of reliability. For the PCK instrument the analysis was less detailed since the initial pilot had already removed a sizeable number of items, and there were fewer constructs that were being measured⁷. The overarching goal of the stability analysis was to make the teacher instruments as efficient as possible, meaning they could capture a sufficient amount of information with an acceptable level of reliability in a manageable (for teachers) length.

The student EGMA assessment instruments were managed by EQID and could not be adapted (or changed) in any way.

Second, the teacher and student data were repeatedly analysed in terms of balance and attrition dynamics.

These two dimensions are closely related, as variation in teacher (between baseline and endline) or student (across grades 1–3) attrition dynamics can result in unbalanced samples, meaning that there are significant differences in the composition of the treatment and control samples that can affect the treatment effects analysis. Imbalance can also result from inefficiencies in the initial 50:50 matching procedure due to limitations in the EMIS data that were available to match the schools. These components were analysed by comparing attrition rates as well as background characteristics of teachers and students who had left the cohort (“attriters”). The most powerful tool is the “selective attrition” analysis function that identifies background characteristics that have significantly different attrition dynamics between treatment and control.

The third analysis area was the estimation of treatment effects, which is the key report outcome that attempts to isolate the impact of SMILE programme participation on the teacher and student outcomes of interest. It is important to be clear that the absence of random assignment means that the treatment effects can only be estimated using statistical modelling applied to a matched design, and these results cannot be assumed to be strictly causal. Nevertheless, with three separate cohorts of teachers (grades 1, 2 and 3), and a large cohort of students with a panel of four measures of mathematics knowledge, the data provide numerous possibilities for comparing treatment and control study subjects. The various methods are reviewed in more detail in the individual reports, and include difference-in-difference (DID), individual matching (for students) and combined (“pooled”) multivariate regression modelling incorporating random and fixed effects.



⁷ Although the PCK data were analysed using item response theory (IRT) methods to generate additional information on item effectiveness and stability.



The fourth set of activities incorporated diagnostic tools to further assess the robustness of the main findings for teachers and students. First, the “attregtest” command in Stata (version 17.0) was used to implement the two attrition tests proposed in Ghanem et al. (2023).⁸ More details are provided in the reports, although the underlying intuition is quite straightforward: attrition can impact the comparisons between treated and control at endline, so the tests use information on the key outcomes (dependent variables) at baseline to assess the degree to which the samples have changed as a result of attrition. This is a version of the selective attrition test used for assessing balance, but it focuses on how the key outcomes of the study are possibly impacted by attrition.

The second diagnostic activity employed the “qqvalues” command from Stata (version 17.0) to obtain adjusted p-values (or “q-values”) that correct for multiple hypothesis testing (see Newson and ALSPAC study team, 2010)⁹. This is often referred to as a “Bonferroni” correction, but the Bonferroni option is only one of multiple corrections that are available in the package. The multiple hypothesis testing bias is a concern for this study because a large set of outcomes are being measured in both the student and teacher samples. When many outcomes are analysed, it

increases the possibility of finding a “false positive” (or “false negative”) result that is defined as a result that does not reflect a “true” difference but is instead the result of random measurement error. The results for the student and teacher treatment effects are adjusted using the Bonferroni method as well as the method proposed by Hochberg, which is not as restrictive.

The final set of analyses extended the multivariate statistical modelling to identify significant predictors of the teacher and student outcomes. This work overlaps with the treatment effects multivariate analysis with an expanded set of independent variables as predictors. For teachers it is important to see how characteristics like education and experience predict outcomes from the teacher survey and PCK test. For students it is also useful to see how student and family background characteristics predict test scores. The key predictors in this expanded multivariate analysis are the teacher survey and PCK test measures in the student achievement estimations. The central hypothesis in this study is that the SMILE programme is impacting teacher capacity in these areas, which in turn is leading to higher levels of student achievement. The statistical modelling makes it possible to test this theory of change.

⁸ Ghanem, Dalia, Hirshleifer, Sarojini, Ortiz-Becerra, Karen. 2023. Testing Attrition Bias in Field Experiments. *Journal of Human Resources*, <http://doi.org/10.3368/jhr.0920-11190R2>.

⁹ Newson, R., and the ALSPAC Study Team. 2010. Software update: st0035_1: Multiple-test procedures and smile plots. *Stata Journal* 10: 691–692.



ស.វ: 39 ឆ.
ស.ស: 19 ឆ.
ឆ.វ: 05

គណិតវិទ្យា មេរៀន

*ប្រភាគចំនាត់ចូលនៃកំរិត



$$\frac{1}{4}$$

$$\frac{2}{4}$$

$$\frac{3}{4}$$

03

MAIN FINDINGS: TEACHERS

Main findings: Teachers

The teacher survey and pedagogical content knowledge (PCK) test findings are summarised based on the five sets of analyses that were reviewed in section 2.6. As noted repeatedly in this discussion, the total amount of output across three years of analysis is extensive, so the summary relies primarily on a description of the main results with individual examples, tables and graphical visuals when necessary.

3.1/ Construct stability

The construct stability analysis for the teacher data incorporated factor analysis and reliability metrics after each data collection round to identify individual items—or blocks of items (i.e. constructs)—that were not generating reliable information. The modifications were concentrated in the pilot and grade 1 teacher data collection periods, and no changes were made in grades 2 or 3. The streamlining of the teacher instruments addressed problems with inefficient items and was also necessary to respond to teacher concerns in the initial data collections that the instruments were too long.

The very detailed summaries of the teacher factor and reliability analysis show a high degree of consistency in the individual factor loadings and reliability measures. This is true for individual questions, specific constructs, and overall measures (such as overall reliability for

the PCK test). Table 1 provides an example of the teacher survey analysis for section 2 of the survey (“Mathematical content and pedagogy”) taken from the grade 2 teacher baseline analysis report.¹⁰ In the first part of the table the grade 1 and 2 baseline scores on the individual items for three constructs (“Number”, “Measurement” and “Geometry”) are compared, and the results do show that the grade 2 teachers reported significantly higher levels of confidence in some of the individual questions (see asterisks). These kinds of differences between teachers by grade level are not unexpected, especially if younger (and less experienced) teachers are assigned the earlier grades. However, the two columns in the right hand side of Table 1 show that the factor analysis results are nearly identical in both survey moments for these three constructs, with very similar individual factor loadings and overall eigenvalues for the three constructs (2a, 2b and 2c).



¹⁰ BN Consult (2024a). “Grade 2 teacher baseline data analysis report.” Phnom Penh.

Table 1. Comparison of G1 and G2 baseline data and factor analysis results for teacher questionnaire Section 2 (Mathematical content and pedagogy)

Item score (mean):			Factor loadings:	
	G1 base	G2 base	G1 base	G2 base
2a. Number				
Q1	3.84	4.09*	0.72	0.72
Q2	4.18	4.44*	0.77	0.72
Q3	3.75	3.87	0.78	0.84
Q4	3.74	3.81	0.78	0.81
Q5	3.30	3.47	0.83	0.76
Q6	3.72	3.95*	0.68	0.69
2a. Eigenvalue			3.47	3.46
2b. Measurement				
Q7	3.52	3.66	0.87	0.87
Q8	3.62	3.84	0.89	0.84
Q9	3.61	3.83	0.91	0.87
Q10	3.60	3.65	0.88	0.83
Q11	3.67	3.77	0.76	0.77
2b. Eigenvalue			3.72	3.50
2c. Geometry				
Q12	3.37	3.58	0.87	0.89
Q13	3.19	3.49*	0.87	0.89
2c. Eigenvalue			1.52	1.60

Source: Grade 2 teacher baseline report

Table A1 in Data Annex A provides an additional summary of the teacher survey and PCK test performance across all three grades (taken from the grade 3 baseline report). The main finding is that the teacher survey and PCK test content areas consistently have eigenvalues that are greater than 1.0 (minimum acceptable level), and the reliability indicator (from grade 3) is greater than 0.80 in almost all of the

teacher survey components. The reliability levels are lower for constructs with very few questions, and the overall PCK test reliability (0.74) is much lower than the overall teacher survey reliability (0.96), but it is still acceptable in comparison with other PCK studies.¹¹

¹¹ For example, the reliability parameter was 0.62 in a 2017 study from Peru using a similar instrument: Cueto, S., León, J., Sorto, M. A., & Miranda, A. (2017). Teachers' pedagogical content knowledge and mathematics achievement of students in Peru. *Educational Studies in Mathematics*, 94(3), 329–345.

3.2/ Attrition and balance comparisons

Figure 1 summarises the teacher attrition rates by treatment status and grade. The attrition rate is defined as the percentage of the original baseline sample that does not have data at endline. The main driver of attrition is “mismatched” teachers that were selected for the baseline but then changed grades during the school year. In all three years the teacher baseline data were collected before the school year actually began

in order to work with treatment teachers before they received the VVOB training. Mismatched teachers were replaced at endline by the actual teacher who was working with the target grade. In other cases the teacher was absent at endline, had transferred to another school, or was a contract teacher at baseline that was no longer employed.

Figure 1. Teacher attrition between baseline and endline by treatment status and grade



Data source: VVOB teacher survey, various years



The attrition rates are similar between treatment and control samples in the grade 1 and 3 surveys, but are significantly different between treatment and control samples in grade 2. Teacher attrition is generally declining by grade, which likely reflects more certain teacher assignment at the beginning of the year in higher grades.

Attrition can impact the analysis of treatment effects by changing the composition of the samples *in different ways between treatment and control*. This idea was tested in all three grades using the selective attrition statistic that identifies individual teacher characteristics that have significantly different attrition dynamics between treatment and control. This was done for a large set of teacher background characteristics, but the main finding across the three years is that there are very few variables that show significant selective attrition. **In the grade 1 and 2 samples only two teacher characteristics (out of 30) showed significant selective attrition, and in the grade 3 sample none of the characteristics were significant.**

Selective attrition is related to balance since the overall matched sample (50:50) design is intended to compare two sets of study subjects that are very similar (except the treatment), but over time the samples may change (i.e. become unbalanced). A related issue is balance at baseline, which is measured three times in the teacher

study since three different cohorts of teachers are analysed. There are two main findings from the baseline (pre-treatment) balance comparisons. **First, in grades 1 and 3 the treatment teachers had significantly lower scores on the teacher survey section 2 questions about “Math content and pedagogy” confidence. And second, in all three survey rounds there are a few background characteristics that are significantly different between treatment and control samples (from a large set of indicators).**

Overall, there are two main takeaways from the balance and attrition analysis. **First, there is some evidence of selective attrition and imbalance, but the scope of this problem is not very serious considering the very large set of characteristics that was considered.¹² And second, this evidence does need to be taken into account in the treatment effects analysis.** This is done by including teacher background characteristics in the multivariate modelling of treatment effects to explicitly control compositional differences between the two samples. The pre-treatment differences in the teacher survey questions (mainly section 2) cannot be addressed explicitly in the analysis, but should be taken into consideration when interpreting the findings from the teacher survey treatment effects analysis.

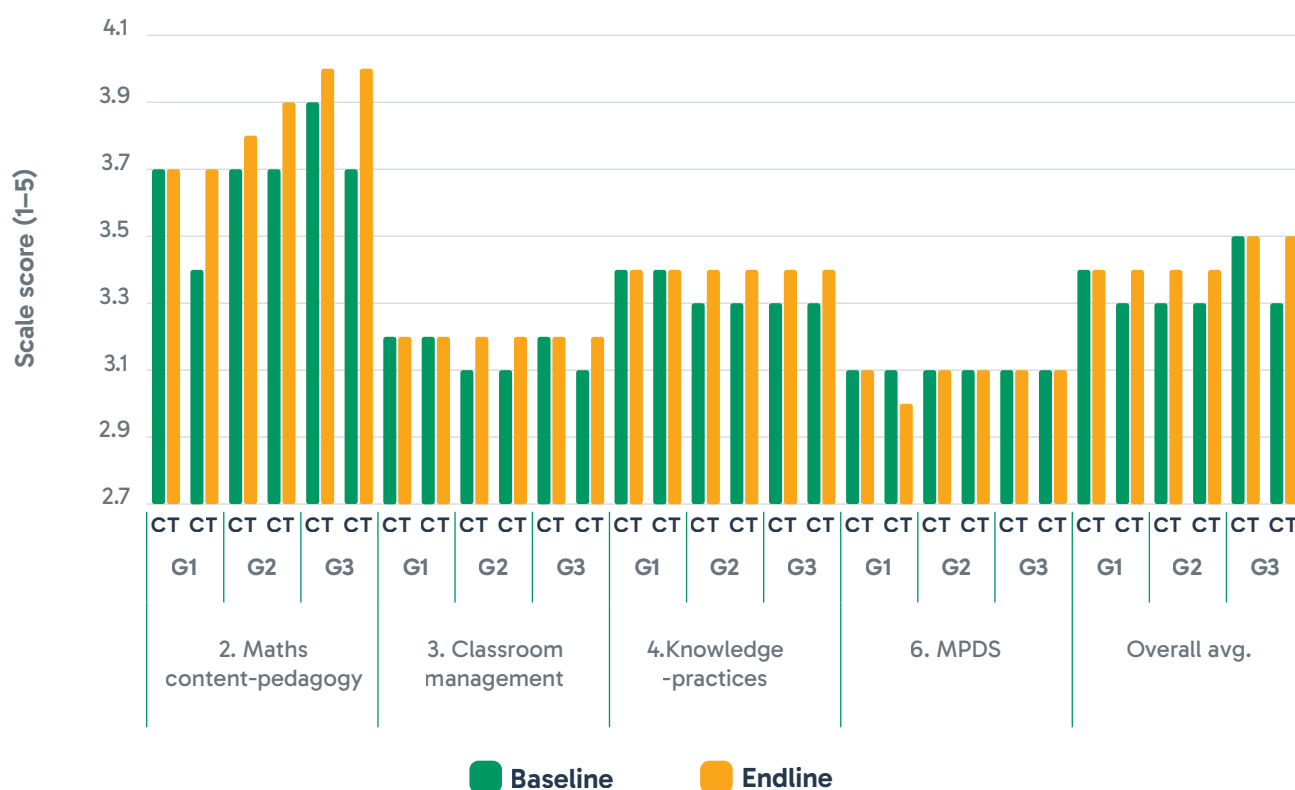
¹² Furthermore, a multiple hypothesis testing correction was not applied to the balance and attrition analyses, despite the very large number of indicators that were considered. This is a more conservative approach that may overstate the degree to which there are significant differences between treatment and control for the attrition and balance comparisons.

3.3/ Treatment effects and diagnostics

The main findings from the treatment effects analysis and diagnostics are summarised together. Figure 2 begins with a basic comparison of treatment (Kampong Thom) versus control (Pursat) teachers for the four sections of the teacher survey and the overall average. The teacher survey questions are all based on a 1-5 ordinal scale. The results show that in most cases

the endline averages are higher than the baseline averages. But VVOB programme teachers (see “T”) have higher endline averages than control teachers (“C”) in only one comparison (survey Section 2, Grade 2). And there is one endline comparison where control teachers have a higher average than treatment (survey Section 6, grade 1).

Figure 2. Comparisons of baseline and endline teacher survey results by grade, treatment status and survey section



Data source: VVOB teacher survey, various years



There is no evidence of VVOB programme impact on teacher PCK.



There is also limited evidence in Figure 2 that treatment teachers experienced larger improvements between baseline and endline on the various teacher survey question blocks. The main exception is in survey section 2 (Maths content-pedagogy) where in all three grades the treatment teachers have higher or equal endline scores despite starting with lower (or equal) scores at baseline. The larger improvements in section 2 for treatment teachers—combined with very similar averages on the other teacher survey sections—result in marginally higher rates of improvement for treatment teachers on the overall teacher survey in all three grades, although again with the caveat that endline scores are virtually identical between control and treated in all comparisons.

There is no evidence of VVOB programme impact on teacher PCK. Figure A1 summarises the overall percentage correct on the PCK test at baseline and endline for each treatment group (by grade). In all three grades the baseline and endline scores are very similar, and the same is true for the net change between baseline and endline (i.e. gain score). The actual changes between baseline and endline are difficult to interpret due to modifications to the test content between baseline and endline, although this is mainly the case for grade 1.

The basic comparisons in Figure 2 (and A1) provide initial estimates of programme impact. Table A2 in the Data Annex (from the grade 3 teacher endline report) provides an example of the more detailed follow-up analysis with comparisons between grade 3 control and treatment at baseline, endline, the difference between baseline and endline, and then three treatment effects estimators based on difference-in-difference and multivariate regression specifications. Additional regression work used pooled data (i.e. merging all three teacher cohorts), different estimation methods (i.e. school random effects) and included all teachers, even those who only had baseline or endline data.

The main findings from Table A2 provide a good summary of the overall findings for the three grades, as well as the results from additional estimations (not presented). **The headline finding is that treatment teachers have significantly larger gains between baseline and endline for the overall teacher survey instrument, as well as the Section 2 and (to a lesser degree) Section 3 content.** The effect sizes are fairly large, with an overall gain score advantage of 0.30–0.44 standard deviations, and an even larger effect for the Section 2 content (0.50–0.60 SD). Furthermore, the treatment effects estimates for the overall survey and the Section 2 difference are mostly significant even when applying the multiple hypothesis testing correction (“q value correction”), and are generally replicated in the other analyses (pooled data, etc.). The same is true with the insignificant results for the PCK test comparisons at the bottom of Table A2: as with the basic comparisons of averages (Figure A2), there is no evidence from the more rigorous treatment effects analyses of a VVOB programme effect on this second teacher instrument.

The evidence of a programme effect for the overall survey and section 2 is a positive result that is generally consistent across all three data collection and analysis periods. However, there are some additional caveats to be noted. First, there is some evidence of non-random attrition patterns that potentially affect the comparisons of these teacher survey indicators at endline (see significant diagnostic parameters in the far right column of Table A2). **Also, the significant programme impact is concentrated in the “Maths content and pedagogy” section—especially “Training application”—and reflects a higher rate of improvement for treatment teachers from a lower baseline average in two of the three teacher cohorts; these more detailed sub-content area treatment effects results (for the grade 3 cohort) are summarised in Table A3.** The fact that treatment teachers reported more improvement in their confidence in geometry, measurement and especially the application of training content is an important finding, and is consistent with programme intent. And this improvement in questionnaire section 2 content is large enough to generate a fairly large treatment effect for the overall teacher questionnaire average. Nevertheless, the bottom line is that the treatment effect for Training Application difference is the only robust result among the various significant differences, and when removing this content area there is much less evidence of overall programme impact.



3.4/ Predictors of teacher outcomes

The most important predictor of the teacher survey and PCK test results is the treatment status measure. But from a more general research perspective it is also important to understand how other teacher characteristics are associated with these indicators. In all three grades the teacher survey and PCK test results were analysed at endline using multivariate analysis. The summary in Table A4 in the Data Annex is taken from the grade 3 report, and is based on a pooled data specification that brings together all three cohorts of teachers into a single estimation.

Overall, four results stand out in Table A4. **First, teacher preparation is significantly associated with the teacher outcomes.** The teacher's years of basic (pre university) education is positively associated with both teacher instrument scores. Teachers with university education score significantly higher on the PCK test, and Contract teachers score significantly lower on both the teacher questionnaire and the PCK test.

Second, female teachers score significantly lower on both teacher instruments, while the results for teacher experience are mixed: more experienced teachers have higher scores on the teacher questionnaire, but lower scores on PCK. It is not unusual for females to score lower in mathematics than males in Cambodia, although in recent national student assessments this is not always the case (see EQID results). For the teacher questionnaire the gender differences may reflect different

levels of confidence, which may not be related to actual teaching ability and preparation. The same argument could apply to the experience result, as older teachers may feel more confident in their overall ability. However, the significantly lower result for PCK for older teachers could reflect differences in pre-service preparation. Entry qualifications for entering the teacher training system have increased over time, and the quality of the training itself has likely improved as well.

Third, there is not much evidence that teacher grade level predicts the teacher outcomes, although grade 3 teachers do have higher overall teacher questionnaire scores than their grade 1 counterparts. It is not clear that teachers in higher primary grades should be expected to have more capacity (or confidence), so this result is not unexpected.

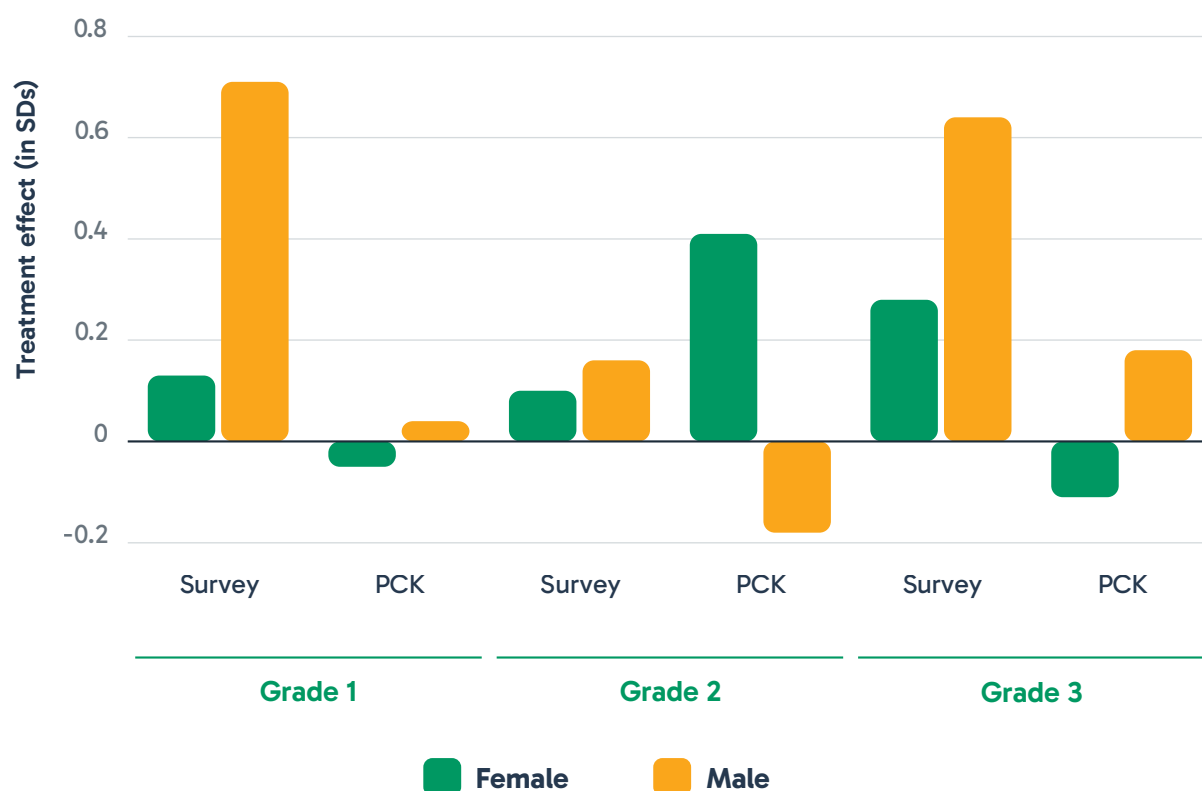
Finally, the teacher questionnaire and PCK test are significantly correlated. The effect sizes for the relationship between the questionnaire and PCK—while significant—are fairly modest: a one standard deviation increase in one measure predicts about a 0.12 standard deviation higher result in the other measure. But this is another form of validation of the instrumentation since it is expected that the teacher's evaluation of their knowledge and confidence across a range of mathematics teaching areas should be related to their "real" knowledge of teaching content.

3.5/ Gender dimensions

The discussion on gender in the various baseline and endline reports focuses mainly on the net difference between female and male teachers in the statistical analysis of the teacher outcomes. As summarised in section 3.4 above, when controlling other factors (including treatment status) males score significantly higher on the teacher survey and the PCK test in the multivariate modelling that brings all three cohorts together. Very little attention was given to differences in the VVOB programme treatment effects by gender, which is a different question.

Figures A2 (teacher survey) and A3 (PCK test) provide a graphical summary of the differences by gender among treatment school teachers. **The results show that male averages are higher than female averages in every comparison, and that they experienced larger gains between baseline and endline in every comparison except grade 2 PCK.** The underlying explanation for the male advantages in the teacher survey and PCK have already been discussed in section 3.4. But the larger gains made by males in Figures A2 and A3 also raise the possibility that they are more positively impacted by exposure to the VVOB training and support programme.

Figure 3. Comparisons of treatment effects by gender, grade and teacher outcome



Data source: VVOB teacher survey, various years

Figure 3 summarises the results from separate treatment effects analyses by gender. The parameters refer to the VVOB treatment effect in standard deviations from the multivariate model 2 estimation (see Table A2 and section 2.5 for more details). **The results confirm that VVOB treatment effects are generally larger for males, especially in grades 1 and 3 for the teacher survey, and grade 3 for PCK.** The female treatment effect is noticeably larger only for grade 2 PCK.

Overall, the gender comparisons show that males consistently performed better than females on the teacher instruments, and male VVOB programme participants realised larger gains than females between baseline and endline in most of the comparisons. For

PCK—where the results are not quite as consistent—this finding is plausibly related to advantages that males have in mathematics content knowledge from early grades, although it should be restated that in recent national assessments in Cambodia girls are performing similar to boys in mathematics. For the teacher survey differences the underlying explanation in overall scores (as well as gains) between baseline and endline are harder to explain, and may simply reflect differences in teacher confidence levels. These comparisons of different forms of teacher capacity—self-reported on the survey, externally tested in PCK—do not automatically equate to *teacher effectiveness*, which is returned to in report section 4 in the student data analysis.

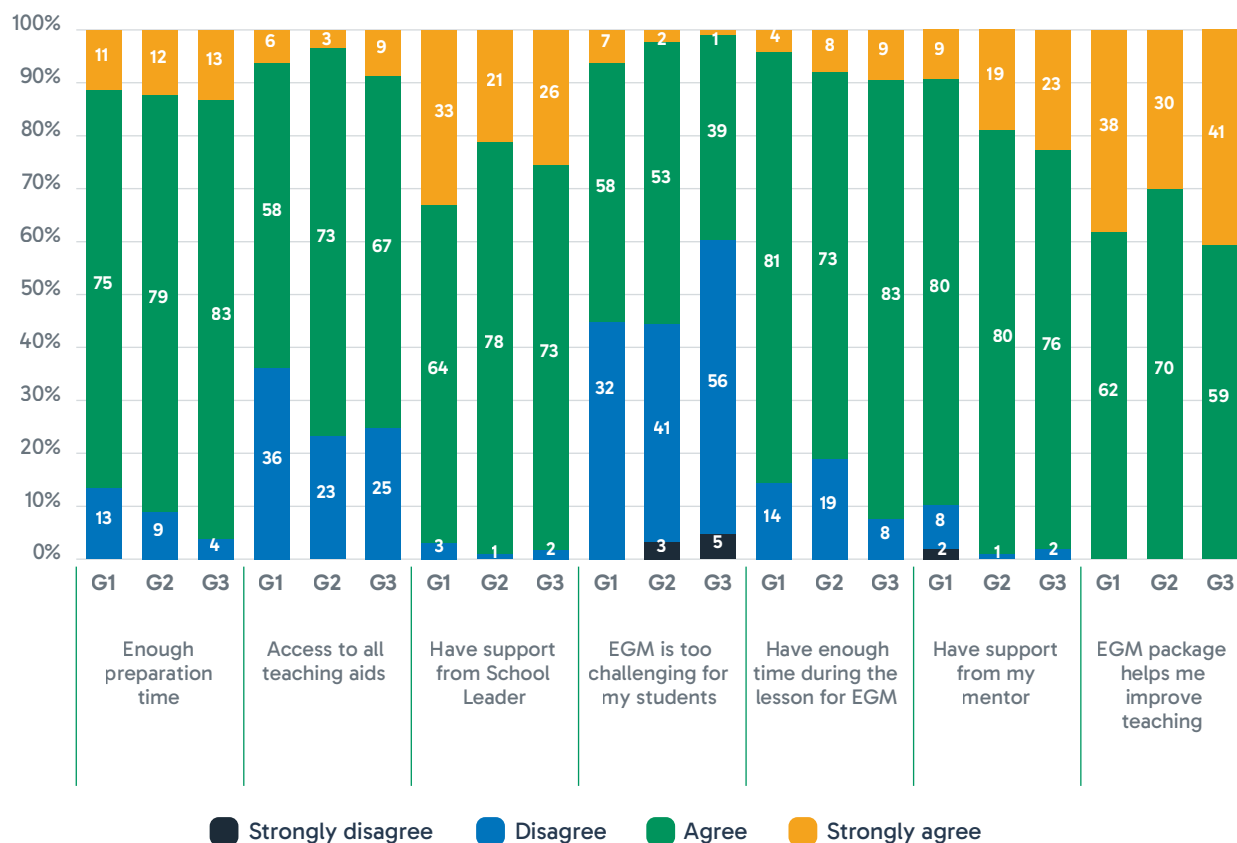
3.6/ Formative feedback

The VVOB teacher survey instrument included blocks of questions for VVOB programme participants to provide feedback about the programme itself. The Figure 4 summary begins with a block of formative feedback questions that were included in all three endline survey rounds. **The main result is that VVOB respondents are generally positive about the training experience, with large proportions of teachers indicating they “Agree” or “Strongly Agree” that they have enough preparation time, have support of their school leaders and mentors, and their teaching is improving as a result of programme participation.**

There are some sizeable shares of teachers who indicated there are problems with getting access to all programme materials, having enough time for the programme, and student readiness for the new content. **Importantly, there is a clear, positive trend in the teacher responses across the three years of data.** This is consistent with teachers becoming more familiar with the EGM packages, and can also be a result of improvements in support functions.



Figure 4. Grade 1, 2 and 3 treatment province teacher responses about VVOB EGM programme participation experiences (“To what extent do you agree with these statements”)



Data source: VVOB endline surveys, 2023–2025

Figure A4 in the Data annex provides the summary of a second set of formative responses from teachers from the grade 2 and 3 surveys about the frequency of engaging in programme activities. These results are more mixed compared with the content in Figure 4. Almost all of the responses are in the “Sometimes” (grey shading) or “Often” (yellow shading) categories (or more frequent), with relatively few teachers indicating “Rarely” or “Never”. The most common activities are “Develop teaching and

learning aids” and “Can use aids in many situations”. The least frequent activities are “Ask students to bring materials” and “Have problems to find materials to make aids”. One result that is somewhat divergent from the results in Figure 4.1 is the grade 2–3 trend: there is no clear evidence that grade 3 teachers are more actively engaging in these aspects of the programme content compared with grade 2 teachers who were working in the same schools and presumably face similar issues.



04

**MAIN FINDINGS:
STUDENTS**

Main findings: Students

The student data collection took place in four moments: grade 1 baseline at the beginning of the 2023 school year, followed by end of year assessments in grades 1 (2023), 2 (2024) and 3 (2025). The analysis summary is similar to teachers, with separate sections for attrition and balance (4.1), treatment effects (with diagnostics) (4.2), predictors of student achievement (4.3) and gender comparisons (4.4). Construct stability is not relevant for students given the reliance on the official Cambodian version of the Early Grade Mathematics Assessment (EGMA) in all four tests (which cannot be modified), and there is also no formative feedback from the student interview on VVOB programme participation.

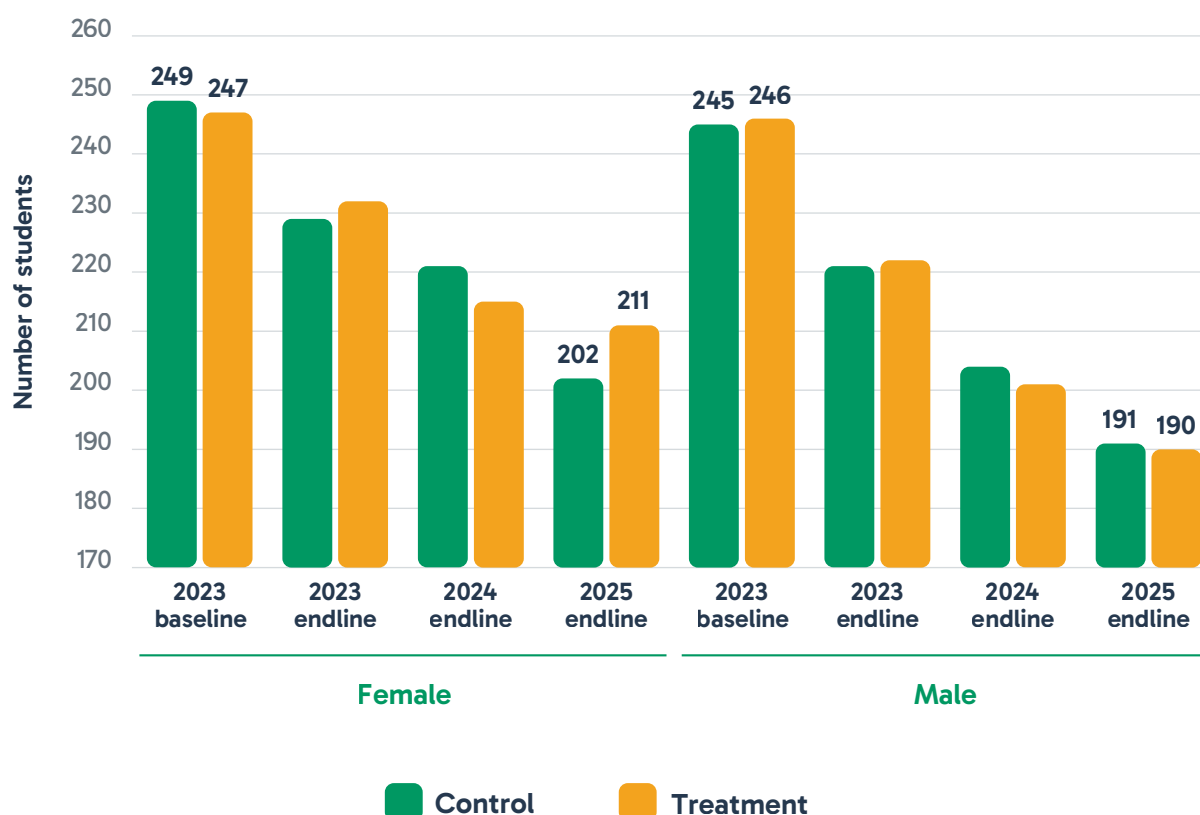
4.1/ Attrition and balance

987 students took the grade 1 baseline EGMA at the beginning of the 2023 school year. The EGMA was repeated at the end of the 2023 school year, and more difficult versions were applied at the end of the 2024 and 2025 school years. **Overall, roughly 20 percent of the original 987 students were unavailable at the time of the 2025 endline assessment.** The individual reports summarise the cohort attrition reasons (dropout, transfer, absent, etc.) by grade. At the time of the 2025 assessment 637 (or 64 percent) of the original cohort was enrolled in grade 3, and 16 percent was in grade 2. The overall attrition rates through the end of 2025 are nearly identical between treatment and control samples (21 percent for control, 19 percent for treatment).

Figure 5 provides a more detailed graphical summary of cohort progress across the four data collection moments by gender and treatment status. The results confirm the very similar cohort trajectories by treatment group, especially for males where the cohort numbers are nearly identical for treatment and control in all four data collection periods. For females there is somewhat more attrition in the control group (19 percent) versus treatment (15 percent). **Also, males are more likely to be missing in all three follow up assessments, and the overall attrition rate at the end of 2025 is significantly higher: 17 percent for females versus 23 percent for males.** It should be re-stated that attrition does not mean the students are no longer in school, but is a combination of dropping out, transferring and not being in attendance on the day of the test.



Figure 5. Summary of student assessment cohort sample sizes by school year, treatment status and gender



Data source: VVOB student surveys, 2023–2025

Several dimensions of sample balance were analysed throughout the four rounds of student surveys. First, student and family background characteristics from the student interview were compared between treatment and control samples after the grade 1 baseline. **These “pre treatment” balance comparisons are critical to establish that the two samples are very similar before the programme intervention began, and the results did confirm that none of the background characteristics were significantly different.** These comparisons also considered student results on the EGMA, which were not significantly different in any of the EGMA test sections.¹³

Balance comparisons were then continuously updated at endline for the three remaining EGMA applications at the end of 2023, 2024 and 2025. These comparisons did find significant differences between students who were no longer in the cohort (“attritors”) and the students who remained, but these attrition-balance dynamics were not significantly different between treatment and control in almost all of the comparisons; **in other words, there was very little evidence of “selective attrition” that can significantly change the composition of one of the samples vis-à-vis the other.** This did vary somewhat by period, as the 2024 endline report does show some significant variation in the attrition dynamics for a couple of background indicators, but the most recent (2025) analysis found no significant variation.

¹³ BN Consult (2023a). “Teacher and student baseline data analysis report.” Phnom Penh.

4.2/ Treatment effects and diagnostics

The use of a single student cohort with as many as four observations per student means there are more analysis options available. Table 2 summarises student percentage correct (and other measures) by EGMA test section for control and treatment at each survey moment. **The main finding is that treatment sample scores are significantly higher than control sample scores in almost all of the comparisons in later stages of data collection when most students were in grades 2 (2024) and 3 (2025).** For example, treatment and

control groups both answered about six percent of Addition 1 content questions correctly in the grade 1 pre-treatment baseline, but by the 2025 endline the treatment student advantage was nearly 10 percent (65% versus 56%). **In standard deviations (“z-score”) the treatment student advantage on the overall EGMA test is 0.30 standard deviations at the end of 2024, and 0.32 SD at the end of 2025.**

Table 2. Student EGMA scores by treatment group, content area and data collection moment

EGMA outcome:	Average by treatment:		p-value
	Control	Treatment	
Number identification:			
Grade 1 baseline (2023)	20.9	23.3	0.33
Grade 1 endline (2023)	67.7	69.1	0.63
2024 endline	87.1	89.1	0.29
2025 endline	93.6	93.9	0.81
Number discrimination:			
Grade 1 baseline (2023)	14.9	14.1	0.69
Grade 1 endline (2023)	61.2	66.5	0.09+
2024 endline	82.2	87.1	0.03*
2025 endline	92.8	93.7	0.52
Addition 1:			
Grade 1 baseline (2023)	5.7	5.6	0.93
Grade 1 endline (2023)	33.9	35.4	0.63
2024 endline	42.6	50.3	0.01*
2025 endline	55.6	64.6	0.01*
Subtraction 1:			
Grade 1 baseline (2023)	3.4	1.8	0.06+
Grade 1 endline (2023)	26.4	27.2	0.81
2024 endline	31.2	37.7	0.02*
2025 endline	42.6	50.9	0.01*
Number pattern			
2024 endline	35.2	41.9	0.04*
2025 endline	48.0	56.0	0.02*

Table 2. Student EGMA scores by treatment group, content area and data collection moment (continued)

EGMA outcome:	Average by treatment:		p-value
	Control	Treatment	
Addition 2			
2024 endline	17.3	24.6	0.02*
2025 endline	33.8	42.5	0.02*
Subtraction 2			
2024 endline	6.5	11.7	0.03*
2025 endline	18.4	24.7	0.09+
Overall percentage correct:			
Grade 1 baseline (2023)	10.7	10.8	0.94
Grade 1 endline (2023)	45.3	47.1	0.50
2024 endline	50.1	55.7	0.01*
2025 endline	61.2	66.9	0.01*
Standardised average ("z-score"):			
Grade 1 baseline (2023)	-0.003	0.003	0.94
Grade 1 endline (2023)	-0.04	0.04	0.50
2024 endline	-0.15	0.15	0.01*
2025 endline	-0.16	0.16	0.01*
IRT equated score:			
Grade 1 baseline (2023)	-1.02	-0.96	0.37
Grade 1 endline (2023)	0.04	0.07	0.51
2024 endline	0.36	0.52	0.01*
2025 endline	0.68	0.90	0.01*

Data source: VVOB G1, G2 and G3 surveys

Notes: *Difference (p-value) is statistically significant at $p \leq 0.05$ level.

+Difference (p-value) is statistically significant at $p \leq 0.10$ level.



The VVOB student advantage is generally larger in more advanced mathematics topics, including the Number Pattern, Addition 2 and Subtraction 2 content areas that were added to the 2024 and 2025 endline EGMA. By contrast, on easier content there is some evidence of programme impact after the initial baseline test, but for Number Identification and Number Discrimination the overall averages were above 90 percent correct by the 2025 endline, so there is not much room for a programme effect.

Table A5 in the Data Annex provides the detailed treatment effects summary by EGMA test section for students with complete grade 1 baseline and 2025 endline data. **The results are generally consistent with the summary in Table 2 above, and show large (and significant) VVOB programme advantages in Addition 1, Subtraction 1 and the overall score.** The effect sizes for the overall EGMA scores are somewhat smaller, especially in the regression-based estimations (Model 1 and 2): for example, the overall standardised difference based on percentage correct is 0.29 standard deviations in the basic DID, and 0.19 SD when using multivariate modelling (see “Model 1” and “Model 2”). For the more advanced content (e.g. Addition 2, Subtraction 2) the before and after difference-in-difference estimation is not effective for measuring programme impact because this content was added in 2024 and there is no clean pre-treatment baseline to measure against.

The main findings for the difference-in-difference approach in Table A5 are generally robust to the different diagnostic corrections. The differences for Addition 1 and Subtraction 1 are significant when applying the q-values correction for multiple hypothesis testing. For the overall EGMA measures the VVOB programme impact is only marginally significant for the percentage correct measure after applying the q-value correction. And there is limited evidence that attrition dynamics are significantly impacting the measurement of programme impact (see far right column in Table A5).

Table 3 concludes the treatment effects analysis with a summary of fixed effects estimation results. The fixed effects modelling is ideal for the data analysis since each study subject has up to four separate observations. The methodology is discussed in more detail in the most recent reports for 2024 and 2025. **The results in Table 3 are similar to the results in Tables 2 and A5, and show relatively large (and significant) treatment effects in Addition 1 and Subtraction 1 and smaller, but significant, effects for overall EGMA test measures.** One limitation with the fixed effects approach is that the new EGMA content (Addition 2, etc.) cannot be analysed since there is no variation in student treatment status for these test sections.

Table 3. Fixed effects estimations of VVOB programme impact by test area

Test outcome	Fixed effects estimator:		
	G1 base+ G3 end only	Data in all four moments	All students
Number identification	-0.08(0.08)	-0.04(0.07)	-0.05(0.07)
Number discrimination	0.04(0.06)	0.13*(0.05)	0.10+(0.05)
Addition 1	0.29*(0.09)	0.20*(0.08)	0.18*(0.07)
Subtraction 1	0.39*(0.11)	0.30*(0.10)	0.24*(0.10)
Overall EGMA score (pct)	0.19*(0.07)	0.16*(0.07)	0.13*(0.06)
Overall EGMA score (IRT)	0.21+(0.012)	0.13(0.10)	0.08(0.10)
Overall EGMA score (IRT global)	0.17*(0.08)	0.10(0.07)	0.05(0.07)
Sample size	1,598	3,016	3,554

Data source: VVOB baseline and endline surveys, 2023–2025

Notes: Standard errors (in parentheses) corrected for clustering of students in schools.

*Parameter is statistically significant at $p \leq 0.05$

+Parameter is statistically significant at $p \leq 0.10$

The main takeaway from the student assessment treatment effects analysis—which is the most important aspect of the VVOB programme evaluation—is that there is clear evidence of programme impact.

The treatment and control students began the treatment period with very similar levels of achievement, but by the end of the testing period the VVOB students were clearly outperforming their control school counterparts. Furthermore, the treatment and control samples have similar attrition dynamics, which enhances the comparability of the samples across all the data collection periods. The effect sizes are relatively large in individual test content areas (upwards of 0.50 SD or more), but the overall EGMA test performance metrics provide the best measures for gauging programme impact. These results are somewhat variable, but the overall programme impact is consistently in the 0.10–0.20 standard deviation range based on the most rigorous treatment effects estimators.



These results are somewhat variable, but the overall programme impact is consistently in the 0.10–0.20 standard deviation range based on the most rigorous treatment effects estimators.



4.3/ Predictors of student test scores

In all four data collections the student test scores were analysed using multivariate methods to identify significant student, family and teacher characteristic predictors. The main finding from this analysis is that very few variables are consistently significant across the different specifications. The main results include:

- Student home possessions (SES), teacher is female, teacher experience, the teacher provides tutoring classes, and multigrade teaching are positively associated with student achievement, although when using student fixed effects specifications almost all of the teacher characteristics are insignificant;
- Student age is a significant negative predictor of achievement, which is probably related to grade repetition;
- The two teacher instrument indicators (teacher questionnaire and PCK test) are generally insignificant. The overall teacher PCK indicator does approach statistical significance in some estimations, but the effect size is very small;
- Even when controlling teacher characteristics the treatment control is generally significant, and of a similar magnitude to other treatment effects analyses (about 0.20 SD for the overall student test score measure).

The general lack of consistently significant predictors of student achievement highlights the inherent difficulty in explaining student achievement differences between students and schools and the complex, multi-factor nature of learning. The lack of robust effects for the teacher instruments is especially important. It was already shown (see report section 3) that the teacher scores on the teacher survey are significantly higher in VVOB treatment schools. However, since the overall teacher survey indicator—as well as the various sub-indicators that are part of the questionnaire—are not significantly associated with student achievement, it is not possible to make any linkage with programme effects on teachers and student outcomes. And it should be restated that there was no evidence of programme impact on teacher PCK.

In sum, the VVOB treatment effect on student achievement is in effect “un-identified” since the underlying programme mechanism that explains the apparent programme impact is unknown. The absence of this linkage is not a serious deficiency in the analysis given the inherent challenges in identifying programme effects on student outcomes and linking those differences with programme treatment mechanisms. Nevertheless, it does raise questions about the efficacy of the teacher instrumentation and whether or not it is truly capturing programme impacts on teachers. For example, the programme may be having an effect on teacher actions in the classroom that are not being captured by the teacher survey or the PCK test. For this reason, VVOB has commissioned an additional study that is collecting data through classroom observations.

4.4/ Gender dimensions

There are three topics where there is potential variation by gender: student achievement treatment effects, predictors of student achievement, and student attrition. At baseline the VVOB programme boys scored marginally higher than girls on the EGMA test (0.06 standard deviations, not significant). At the time of the final endline (2025) their scores were virtually identical, so based on simple averages there is not much evidence that the VVOB programme impact varied by gender. As a reference point, in the control schools the baseline male advantage was about 0.03 standard deviations, and in the 2025 endline the girls actually scored about 0.10 SD higher, but the difference is not statistically significant.

Table A6 in the Data Annex summarises the fixed effects parameter results for VVOB programme treatment effects on EGMA scores by gender. The programme effect is larger for boys in Addition 1, Subtraction 1 and the overall test average. However, when analysed separately using interaction terms there is no evidence of significant treatment effects variation by gender.

For student achievement the main takeaway is that VVOB programme exposure has no significant effect on the gender achievement gap, which in this sample was not very large at the beginning (or end) of the testing period. In both the treatment and control samples the girls are catching up with boys on the overall EGMA test measure over time, but there are no significant differences by gender in these dynamics overall or within the VVOB sample.

Section 4.3 summarised the predictors of student achievement by gender. The same models were estimated separately by gender, but there are very few notable differences. Student age (negative predictor) and SES (positive) are more significant in the boys-only sample. Another gender-related result for achievement is that student scores are significantly higher in some estimations when they are studying with female teachers. However, there is no evidence that gender matching is important (i.e. girls do better with female teachers), and this effect does not significantly vary between treated and control samples.

Finally, when student attrition among the baseline cohort (all 987 students) is modelled there are very few significant predictors, or differences by gender. As noted above, student attrition can result from dropout as well as transferring and being absent on the day of the test. The student's initial test scores, family background characteristics (like SES), teacher characteristics and the VVOB programme control are not significantly related with the probability of not continuing in the sample. The only significant predictor is age, as older children at baseline are less likely to be present in the study

sample in later moments. For girls the only significant predictor of attrition is age (older girls at baseline more likely to be missing later on). For boys there are several indicators that are significant (or marginally significant) predictors of attrition: age (positive), home possessions (SES, negative) and multigrade schools (positive).

“

Finally, when student attrition among the baseline cohort (all 987 students) is modelled there are very few significant predictors, or differences by gender.

”





05

**VVOB PROGRAMME
EFFECTIVENESS
DISCUSSION**

VVOB programme effectiveness discussion

This section provides a brief review of relevant evidence, as well as basic cost-effectiveness calculations (and comparisons), to assess the VVOB evidence in a larger national and international research context.

5.1/ Recent evidence from Cambodia

For national evidence there are only a handful of studies during the Early Grade Learning (EGL) period in Cambodia (starting around 2018) that provide relevant evidence for the VVOB programme. Not only is this a small literature base, but the studies tend to focus on EGRA rather than EGMA, the samples in some cases are quite small, the designs vary in terms of the degree to which causality can be inferred, and the studies do not report effect sizes using the same standardised approach incorporated in the VVOB study.

Despite these limitations in terms of comparability, the most important takeaway from the existing Cambodia evidence is that the EGL interventions are consistently associated with moderate improvements in student achievement. An evaluation of the Easy to Learn (E2L) project in Kratie province found significantly higher EGRA scores for students in treatment schools during the 2019–2021 period, with standardised effect sizes of 0.10–0.30 SD.¹⁴ USAID’s Inclusive Primary Education Activity (IPEA) found significant difference-in-difference (DID) treatment effects in several EGRA sections between a 2021 baseline and 2022 endline for treatment provinces compared with a control province (Pursat), although the differences were not very large (standardised effects not provided).¹⁵ The STEPCam project collected EGRA and EGMA data on students in

a 2019 baseline, followed by an endline data collection that took place during a shortened school year (due to Covid-19). The results show general declines in student scores, but student scores in the Early Grade Learning (EGL) intervention provinces declined at a lower rate in both subjects.¹⁶ Finally, the Cambodia Development Partnership Fund (CPDF) early grade learning initiative was evaluated using EGRA and EGMA tests in grade 1 in a pre-treatment baseline (2022) with follow-up in 2023 and 2024. The results show moderate gains in EGRA, and somewhat larger gains in EGMA at the time of the first follow-up (2023). And then larger gains in EGRA (EGMA results not presented) in the 2024 follow up.¹⁷ Effect sizes are not provided or calculable.

A very general reference point for Cambodian student performance in recent years is provided by the SEA-PLM 2019–2024 regional assessment. **Cambodian student scores in 2024 showed significant improvement against the 2019 baseline, with roughly 0.30 SD improvement in reading and an even larger (0.60 SD) improvement in mathematics.** These improvements are smaller when controlling differences in student background, but the main finding still stands out: there is evidence of significant student achievement improvement during the EGL period in Cambodia.

¹⁴ Chhorn Chhayyuth (2023). *Annual EGRA Report: Easy to Learn (E2L) Phase II Project in Chhlong District, Kratie Province*. These effect sizes are estimated here based on the data provided in the report.

¹⁵ RTI International (2024). *USAID Cambodia Inclusive Primary Education Activity (IPEA): Student performance evaluation to guide the scaling of Early Grade Reading*.

¹⁶ EQAD, GPE and UNESCO (2023). *STEPCam Endline EGR and EGM Assessment (power point)*.

¹⁷ EQAD and CPDF (2024, 2025). *Summaries of EGRA and EGMA test results (power points)*.

5.2/ International evidence

International reference points exist but also have limitations due to differences in programme content, study design and reporting metrics. Evidence of the impact of professional development (PD) focused on primary-grade mathematics has been documented in several research studies, including those conducted in low- and middle-income countries. **Average effect sizes reported in meta-analyses focusing on student outcomes are 0.20 to 0.35 SD for mathematics-focused PD, indicating moderate effects that are similar to what was found in this study.**^{18,19,20} These effects are typically smaller than gains in teacher knowledge, reflecting the time required for instructional improvements to translate into student learning. In developing-country contexts, findings are similar but more variable: studies of structured PD models often show effects in the ~0.10–0.30 SD range, with stronger impacts when implementation is high quality and aligned with curriculum and assessment.

Studies that directly measure the impact of training on teacher knowledge (e.g., content knowledge or pedagogical content knowledge) often show larger effects than what is shown for teacher training on student outcomes, sometimes in the ~0.50–0.70 SD range when PD is tightly focused on subject matter and how students learn it.^{21,22} **Evidence from developing-country interventions (e.g., structured coaching or lesson study models) aligns with this pattern: gains are strongest when PD is sustained, content-focused, and practice-embedded, though variability is high due to differences in implementation quality and baseline teacher preparation.**

For teacher self-efficacy—which was measured in this study through the teacher survey—the evidence suggests smaller and less consistent effects, especially in mathematics. Meta-analytic findings indicate that math-focused programmes tend to yield small effect sizes on self-efficacy (generally ~0.10–0.30 SD), lower than those observed in broader STEM or science-focused PD.^{23,24} This is particularly relevant in developing countries, where teachers often begin with lower baseline confidence in math instruction; PD can improve efficacy, but changes are typically incremental unless programmes include coaching, collaboration, and opportunities for mastery experiences. **Overall, the literature suggests moderate gains in teacher knowledge but only modest improvements in self-efficacy, indicating that building confidence may require more intensive or longer-term professional learning designs.**

Programme design plays a central role in determining impact. PD that is content-focused, includes active learning, provides ongoing coaching, and is embedded in teachers' daily work tends to yield the largest gains.²⁵ Short, one-off workshops are generally ineffective. In developing countries, structural constraints such as large class sizes, limited instructional materials, and gaps in teacher preparation can dampen the translation of improved teaching into student achievement. Overall, PD is a reliable but moderate lever for improving student learning, with stronger effects when combined with aligned curriculum, materials, and sustained instructional support.

¹⁸ Kraft, M. A., Blazar, D., & Hogan, D. (2018). The effect of teacher coaching on instruction and achievement: A meta-analysis of the causal evidence. *Review of Educational Research*, 88(4), 547–588. <https://doi.org/10.3102/0034654318759268>

¹⁹ Lynch, K., Hill, H. C., Gonzalez, K. E., & Pollard, C. (2019). Strengthening the research base that informs STEM instructional improvement efforts: A meta-analysis. *Educational Evaluation and Policy Analysis*, 41(3), 260–293. <https://doi.org/10.3102/0162373719849044>

²⁰ Popova, A., Evans, D. K., Breeding, M. E., & Arancibia, V. (2022). Teacher professional development around the world: The gap between evidence and practice. *The World Bank Research Observer*, 37(1), 107–136. <https://doi.org/10.1093/wbro/lkab006>

²¹ Darling-Hammond, L., Hyler, M. E., & Gardner, M. (2017). *Effective teacher professional development*. Learning Policy Institute.

²² Timperley, H., Wilson, A., Barrar, H., & Fung, I. (2007). *Teacher professional learning and development: Best evidence synthesis iteration*. New Zealand Ministry of Education.

²³ Lynch et al., (2019).

²⁴ Sims, S., & Fletcher-Wood, H. (2021). Identifying the characteristics of effective teacher professional development: A critical review. *School Effectiveness and School Improvement*, 32(1), 47–63. <https://doi.org/10.1080/09243453.2020.1772841>

²⁵ Darling-Hammond et al., (2017); Kraft et al., (2018).

5.3/ Cost-effectiveness analysis (CEA)

The initial cost effectiveness analysis (CEA) suggests that the VVOB programme compares somewhat favourably with other interventions (international evidence base), but the results are quite sensitive to model specification and calculation inputs. There are three main inputs for the CEA calculations. VVOB has provided a detailed summary of programme “ingredient” costs for the programme period (2022–2025). This mainly refers to training and orientations for teachers, mentors, school principals and other programme participants, as well as additional staff support and printing costs. The costs for Kampong Thom were estimated on the basis of its share of the total budget across three programme provinces.

The second input is the total number of beneficiaries. This number is quite large since grade 1, 2 and 3 students are the ultimate beneficiaries who are impacted by trained/ supported teachers, and every school in Kampong Thom took part in the programme. For example, there are three cohorts of grade 1 students who have benefitted from the programme (2022–23, 2023–24 and 2024–25), although this assumes that a trained teacher is working with grade 1 students in all three years. For grade 2 there are two cohorts of students who are in VVOB-supported classrooms (2023–24 and 2024–25), and for grade 3 just the 2024–25 cohort. Summing the total Kampong Thom enrolment (from EMIS) for these grades and years, a total of roughly 120,000 students participated in the programme during the three-year study period.

The final input is the return to programme participation (i.e. treatment effect), which has been the critical data point across the three years of data analysis and report preparation. This is another complicated aspect of the CEA because of the reliance on a “triangulated” approach to measuring treatment effects on student achievement, with difference-in-difference (DD) regressions as well as fixed effects models. For the CEA two sets of treatment effects estimates are used: the basic DD parameter comparing the most recent test score against the grade 1 (pre-treatment) baseline, and the fixed effects (FE) estimator for students with complete data. The DD treatment effects are grade specific, while the FE estimates apply to the entire programme participation period. Each estimate has a point estimate as well as an upper and lower bound based on the standard error.

Table 4 presents the CEA results based on two cost effectiveness parameters: the per-student cost to achieve a 0.10 SD improvement in learning, and the change in total standard deviations of learning for each \$100 USD spent by the programme. For the DD regression input the average treatment effect across the three years of analysis (and grades) was 0.18 SD, but a grade-specific parameter was used for the CEA calculations; the grade 1 DD effect was small (and not significant), but it increased for grades 2 and 3. To get the programme costs for improving student achievement by 0.10 SD the average per student costs for the VVOB programme (\$9.01) is multiplied by 0.10 divided by the average treatment effect (0.18). **Using the point estimate values from the treatment effects analysis this comes to \$5.15 (per student) for each 0.10 SD improvement in achievement, and \$8.79 per student based on the FE treatment effects parameters that are smaller than the DD parameters.**

Based on the point estimate average the VVOB programme produces 1.9 standard deviations of learning for each \$100 that is spent using the DD parameters, and 1.1 SDs using the FE parameters.



Table 4. Summary of Cost Effectiveness Ratios (CER) by measure and input parameters

CER measure/parameter:	Cost Effectiveness Ratio (CER):		
	Point estimate	Lower bound	Upper bound
From DD regression:			
Cost per pupil/0.10 SD gain	\$5.15	-\$62.92	\$2.47
Additional SD per \$100	1.9	-0.2	4.0
From Fixed Effects (FE):			
Cost per pupil/0.10 SD gain	\$8.79	-\$16.36	\$3.47
Additional SD per \$100	1.1	-0.6	2.9

Table 4 includes the upper and lower bounds that are based on the 90 percent confidence interval for each point estimate. For example, the grade 3 endline fixed effects estimate for programme impact was 0.16 SD with a standard error of 0.07, which translated to a lower bound of just 0.04 SD and an upper bound of 0.28 SD. The lower and upper bound estimates in Table 4 indicate considerable sensitivity in the CER estimations. Some parameters are actually negative because of the small treatment effects that were found at the end of grade 1, which actually produces a negative parameter in the lower bound which was then applied to all three cohorts of grade 1 students (the largest group). By contrast, if the “true” programme effect on learning is closer to the upper bound estimation the corresponding programme costs are lower for each 0.10 SD produced (\$2.47 for DD, \$3.47 for FE), and more learning is produced for each \$100 of programme spending (4.0 for DD, 2.9 for FE).

These initial estimates for VVOB Cambodia are similar to the CEA results from a government-led TaRL programme in Zambia that reported \$6.38 per 0.10 SD improvement based on a comprehensive mathematics assessment, and more favorable cost effectiveness when focusing just on foundational skills²⁶. Figure 6 provides additional comparisons from JPAL using the \$100 spending metric (see right side of figure). The results show that many of the programmes are in the 1.0–2.0 SD range, but with some very large exceptions.



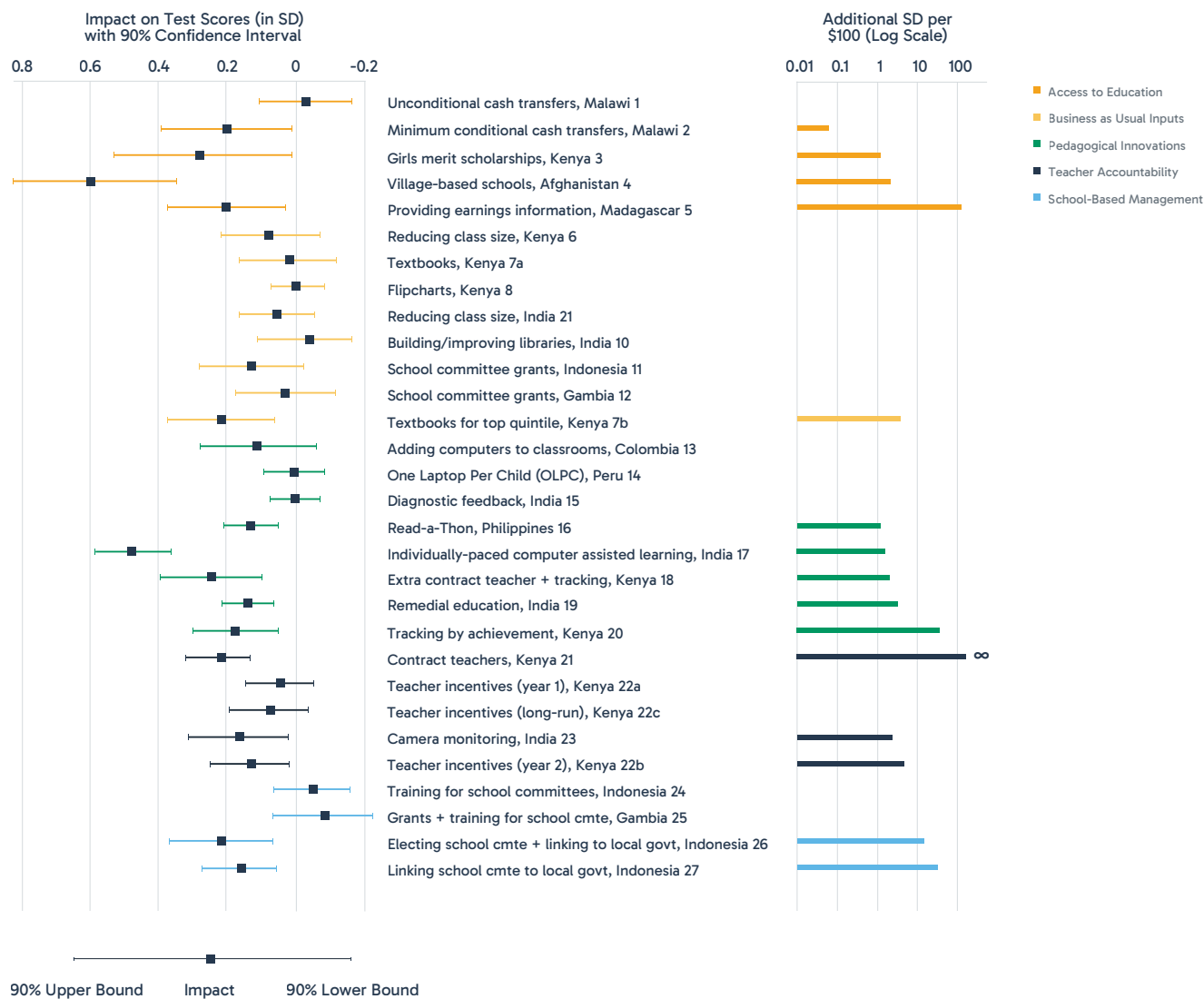
The lower and upper bound estimates in Table 4 indicate considerable sensitivity in the CER estimations.



²⁶. De Barros, A. and T. Lubozha (2026). *Targeting foundational skills at scale: Skill specificity and transfer*. CES IFO Working Papers 12542.

Figure 6. Cost Effectiveness comparisons (from JPAL)

Improving Student Learning: Cost Effectiveness of Education Programmes





06

MAIN FINDINGS AND RECOMMENDATIONS

Main findings and recommendations

The cumulative results from three years of data collection and analysis support two main findings. First, there is substantial evidence of a significant VVOB programme impact on student achievement in Mathematics. This evidence rests on three sets of findings: 1) the treatment and control samples are well balanced at baseline and endline, and there is no evidence of selective attrition for the cohort with complete data at the grade 3 endline; 2) the treatment effects are significant across a range of EGMA content areas, as well overall EGMA score measures; and 3) the treatment effects estimates are generally robust to multiple hypothesis testing adjustments, attrition effects and choice of estimator (difference-in-difference, fixed effects, etc.). In standard deviations the effect sizes are in the 0.10–0.30 standard deviations (SD) range for the overall EGMA test (with larger effects in specific EGMA content areas), which is a small to modest impact.

The second main finding is that there is limited evidence of a programme impact on the teacher efficacy survey content, and no evidence of impact on teacher pedagogical content knowledge (PCK).

The treatment effects for the teacher survey are concentrated in the Section 2 set of questions (“Maths content and pedagogy”), and are especially large for the “Training application” sub-set of questions. These findings are consistent with recent participation in a training programme that includes support for teachers to apply new techniques in the classroom. There is some evidence of programme impact in other areas of the questionnaire, but those results are less robust. The limited treatment effects on the teacher questionnaire content, and the lack of programme effect on the PCK construct, raise questions about the utility of the two instruments and the degree to which they can capture programme impacts. The results are also a reminder of the inherent difficulty in capturing programme effects on teacher attitudes, capacity and behaviours through surveys. This is one reason why VVOB has commissioned a parallel study that is using classroom observations to measure programme impact.

The evidence of a VVOB programme impact on student achievement is encouraging, but the source of this apparent treatment effect is not clear. Neither of the two main teacher indicators (self efficacy survey and PCK) are consistently significant in the modelling of student achievement, so even the modest programme impact on the teacher survey content cannot be directly linked with student outcomes. In other words, the underlying mechanisms that have led to the apparent programme impact on student achievement are not identified through the data collection. That work remains for additional evaluations to address, such as the classroom observation study that can follow up on the intended impact mechanisms that include practice-

level adjustments, the use of structured materials and mentoring support. Additional qualitative data could also be collected from programme participants at all levels (i.e. teachers, directors, district officers, etc.) that unpack these programme components to understand more about the implementation challenges and successes.

These findings are consistent with recent research in Cambodia linking the Early Grade Learning (EGL) package—which the VVOB mathematics intervention is part of—with improvements in student learning.

Most of this evidence is on EGRA rather than EGMA, and the comparability of the designs and reporting is somewhat limited. But the important point is simply that the VVOB evidence is clearly contributing to a growing literature on EGL dynamics in Cambodia, which is critical for continuing the work of improving student learning outcomes through more prepared and supported teachers. Furthermore, it is important to note that the evidence provided here is particularly robust in comparison with most of the other studies, and includes a longitudinal design with a matched set of schools and linked teacher-student data.

The main findings are also in line with international evidence, although these comparisons are even more problematic given differences in interventions, research designs and contexts.

The main takeaways from the international literature are clearly relevant for the Cambodian teacher training context. Teacher professional development interventions have been linked with student achievement gains, but they tend to be fairly modest and are larger when the intervention is content-focused, includes active learning, provides ongoing coaching, and is embedded in the teacher’s daily work. The VVOB programme has some of these features, but there are clear challenges in delivering high quality training and support through the Training of Trainers (ToT) approach that relies heavily on government counterparts.

The initial cost effectiveness analysis (CEA) suggests that the VVOB programme compares somewhat favourably with other interventions (international evidence base), but the results are quite sensitive to model specification and calculation inputs. The costs to improve student achievement by 0.10 SD are between \$5.15 and \$8.79, which is similar to the CEA results from a government-led TaRL programme in Zambia (\$6.38 per 0.10 SD²⁷). Another set of CEA parameters finds that the VVOB programme produces between 1.1 and 1.9 standard deviations of learning for each \$100 of programme spending. JPAL has shown that quite a few programmes globally are in the 1.0–2.0 SD range, but with some very large exceptions (larger and smaller).

Overall, the findings by gender are somewhat mixed. Girls and boys began the EGMA assessment in grade 1 at similar levels (boys were slightly higher), and their scores in the 2025 endline were also very similar. Attrition rates were higher for male students than females across the three-year data collection cohort. Male teachers consistently scored higher on the PCK test and reported higher levels of self-efficacy at baseline and endline, but the programme effects did not vary much by gender. Student test scores are significantly higher when studying with female teachers in some estimations.

The teacher survey also included blocks of questions to provide formative feedback on VVOB implementation from the teacher's perspective. **The main result is that VVOB respondents are generally positive about the training experience, with large proportions of teachers indicating their teaching is improving as a result of programme participation.** Importantly, there is a clear, positive trend in the teacher responses across the three years of data.

“**Student test scores are significantly higher when studying with female teachers in some estimations.**”



²⁷ De Barros, A. and T. Lubozha (2026). Targeting foundational skills at scale: Skill specificity and transfer. CES IFO Working Papers 12542.



The evidence from the teacher and student survey analysis does not provide much insight for formulating recommendations for the VVOB programme implementation specifically, or Early Grade Learning (EGL) initiatives in Cambodia more generally. The main recommendations instead draw more broadly from the PD literature and lessons learned internationally:

- **Programmes that support teacher learning should focus on understanding the big ideas of elementary mathematics content, their representations, and connections. Pedagogical strategies should be actively modelled by showing teachers how to implement them in the classroom. Although challenging to implement, teachers should be familiar with students' ways of thinking and be able to adapt their practice accordingly.**
- **Programmes should provide high-quality materials for students, along with corresponding teacher guides. The support, in the form of mentors or coaches, should be ongoing throughout the school year.**
- **Teachers should also be supported by mentors and/or peers, before and after lessons, by discussing planning and reflection, that is, what went well and what aspects can be improved.**

Teacher feedback on the VVOB programme (through the teacher survey) highlights the importance of several features of effective professional development programmes, and work that remains to improve these areas. Sizeable shares of teachers indicated issues with getting access to all programme materials, having enough time for the programme, and student readiness for the new content. The teacher feedback is consistent, to some degree, with VVOB staff feedback that confirms early implementation challenges (e.g. limited mentoring) and later system strengthening and improved support structures.

Finally, it is important to continue to collect formative feedback data on EGL programme interventions, undertake rigorous assessments of student achievement to isolate programme impact, and also collect systematic data on classroom processes through observations and other modalities. The VVOB impact evaluation study is contributing to a growing evidence base on EGL effectiveness in Cambodia, but there are ways to improve this work going forward. Randomised designs are difficult to implement given the district- or provincial-wide implementation and teacher training and support regimes that are in place, but randomly assigning districts to new treatments provides a plausible way forward for establishing programme impact. Using baseline (pre-treatment) and follow-up assessments (midline, endline) is also critical, ideally using standardised testing instruments (EGRA, EGMA). Results should be reported using standardised metrics (standard deviations), ideally based on rigorous statistical frameworks that include additional controls to sharpen the analysis of programme impact.

Report bibliography

BN Consult (2026). "Grade 3 teacher and student endline report." Phnom Penh.

BN Consult (2025a). "Grade 3 teacher baseline data analysis report." Phnom Penh.

BN Consult (2025b). "Grade 2 teacher and student endline report." Phnom Penh.

BN Consult (2024a). "Grade 2 teacher baseline data analysis report." Phnom Penh.

BN Consult (2024b). "Teacher and student Grade 1 endline report." Phnom Penh.

BN Consult (2023a). "Teacher and student baseline data analysis report." Phnom Penh.

BN Consult (2023b). "Pilot and validation report." Phnom Penh.

BN Consult (2022). "Teacher questionnaire literature review." Phnom Penh.

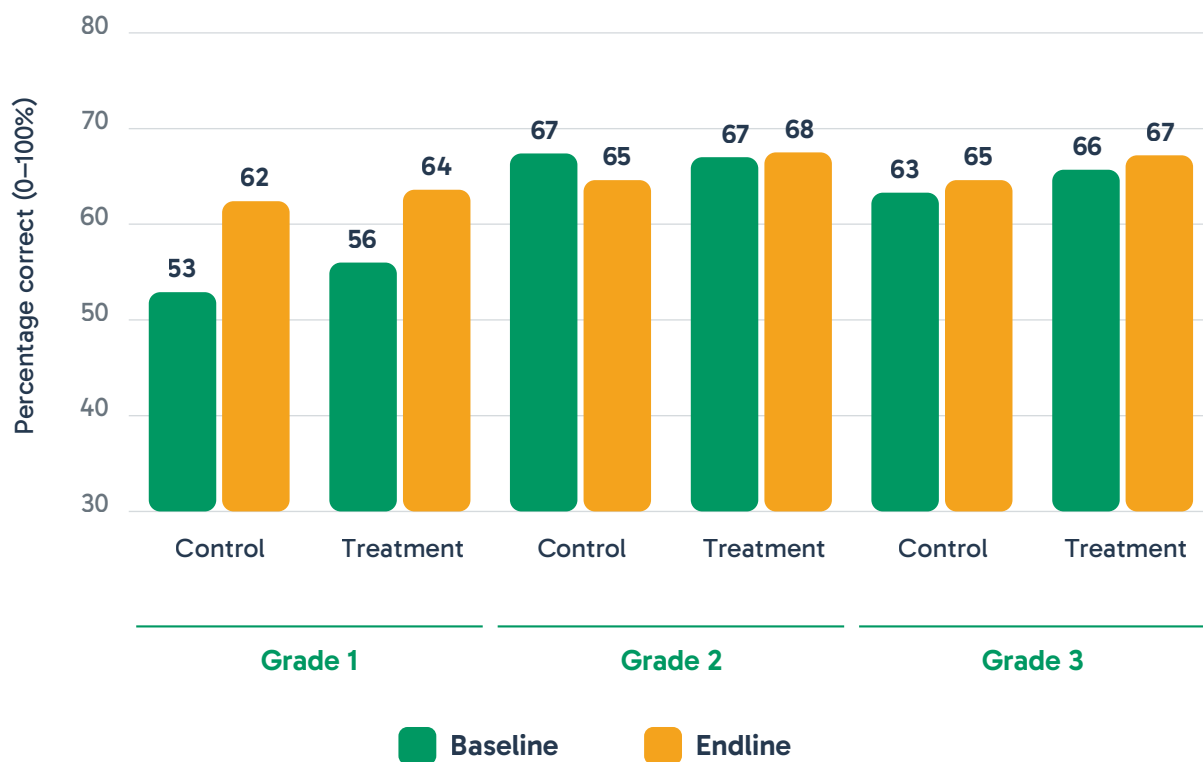
Data Annex A

Table A1. Grade 3 baseline teacher questionnaire detailed analysis

Questionnaire area:	Eigenvalue by baseline:			Reliability G3 (α)
	G1	G2	G3	
Teacher Questionnaire:				
Section 2: Math content and pedagogy				
2a. Numbers (6)	3.47	3.46	3.17	0.87
2b. Measurement (5)	3.72	3.50	3.33	0.91
2c. Geometry (2)	1.52	1.60	1.51	0.58
2d. Learning from trainings	2.08	2.58	3.09	0.94
Overall (17)	8.83	8.51	8.03	0.93
Section 3: Classroom management				
3a. Positive discipline (3)	1.51	1.51	1.41	0.76
3b. Differentiation (5)	1.90	1.73	1.65	0.70
3c. Flexible grouping	2.00	1.64	1.44	0.71
Overall (12)	4.73	4.28	3.95	0.84
Section 4: Knowledge, attitudes and practices				
4a. Strength of beliefs (22)	12.92	9.80	10.48	0.95
4b. Attributes (11)	3.73	2.99	3.34	0.77
4c. Maths anxiety (8)	4.32	3.12	3.00	0.81
Overall (41)	15.33	12.51	12.54	0.94
Section 6: MPDS				
6a. Affect (3)	1.73	1.31	1.80	0.82
6e. Risk taking (1)	----	----	----	----
6f. Goals (3)	1.62	1.30	1.65	0.82
6g. Motivation (1)	----	----	----	----
6h. Self-efficacy (3)	1.79	1.66	1.78	0.81
Overall (11)	7.42	4.30	5.36	0.91
Overall teacher questionnaire (81)	27.76	19.8	20.1	0.96
Teacher PCK Test:				
Numbers and operations (8)	3.14	3.50	3.86	0.72
Fractions and number patterns (5)	2.11	2.27	2.00	0.38
Geometry and measurement (6)	1.94	2.81	2.19	0.43
Problem solving (2)	1.40	1.45	1.28	0.27
Overall (21)	4.83	5.66	3.10	0.74

Data source: VVOB teacher G1 and G2 baseline data collections, 2022–2023

Figure A1. Comparisons of baseline and endline teacher PCK test percentage correct by grade and treatment status



Data source: VVOB teacher survey, various years

Table A2. Summary of grade 3 teacher questionnaire and PCK test results for teachers with complete baseline and endline data

Teacher outcome:	1. Baseline results:		2. Endline results:		3. Difference (SD) (2minus 1):		4. Treatment effects:			Model 1 q-value correction:		Attrition null hypothesis (p-values)
	Control (99)	Treated (88)	Control (99)	Treated(88)	3a. Control	3b. Treated	DD	Model 1	Model 2	Bonf.	Hoch.	
Teacher questionnaire:												
Section 2 (avg.)	3.9(0.06)	3.7*(0.07)	4.0(0.05)	4.0(0.06)	-0.32(0.10)	0.28*(0.10)	0.59*(0.14)	0.56*(0.13)	0.59*(0.14)	0.00*	0.00*	0.03*/0.07+
Section 2 (factor)	0.08(0.11)	-0.31*(0.13)	0.11(0.09)	0.18(0.10)	-0.26(0.10)	0.24*(0.11)	0.50*(0.15)	0.48*(0.14)	0.50*(0.14)	0.01*	0.01*	0.07+/0.14
Section 3 (avg.)	3.2(0.04)	3.1*(0.05)	3.2(0.04)	3.2(0.04)	-0.14(0.10)	0.13+(0.11)	0.27+(0.16)	0.29+(0.17)	0.31+(0.17)	0.55	0.36	0.02*/0.05*
Section 3 (factor)	0.17(0.09)	-0.20*(0.11)	0.11(0.10)	-0.02(0.09)	-0.13(0.10)	0.11(0.11)	0.24(0.15)	0.25(0.17)	0.28+(0.17)	0.79	0.45	0.04*/0.10+
Section 4 (avg.)	3.4(0.03)	3.3(0.04)	3.4(0.04)	3.4(0.03)	-0.10(0.11)	0.09(0.11)	0.18(0.16)	0.17(0.15)	0.19(0.16)	----	----	0.46/0.66
Section 4 (factor)	-0.03(0.10)	-0.22(0.12)	0.09(0.13)	0.11(0.10)	-0.12(0.12)	0.11(0.10)	0.23(0.16)	0.20(0.14)	0.22(0.14)	----	----	0.19/0.18
Section 6 (avg.)	3.1(0.04)	3.1(0.04)	3.1(0.03)	3.1(0.03)	-0.08(0.13)	0.07(0.13)	0.15(0.18)	0.15(0.18)	0.09(0.18)	----	----	0.57/0.33
Section 6 (factor)	-0.01(0.11)	-0.11(0.11)	-0.02(0.10)	0.05(0.09)	-0.09(0.11)	0.08(0.12)	0.17(0.18)	0.18(0.19)	0.11(0.18)	----	----	0.71/0.38
	3.5(0.03)	3.3*(0.04)	3.5(0.04)	3.5(0.03)	-0.23(0.10)	0.21*(0.12)	0.44*(0.16)	0.38*(0.14)	0.39*(0.14)	0.05*	0.05*	0.09+/0.12
	0.02(0.10)	-0.27+(0.12)	0.10(0.13)	0.12(0.10)	-0.21(0.10)	0.19*(0.11)	0.39*(0.17)	0.31*(0.13)	0.32*(0.13)	0.12	0.10+	0.17/0.11
PCK test percent correct	63.3(2.3)	65.7(1.7)	64.6(2.3)	67.2(2.1)	0.004(0.11)	-0.004(0.11)	-0.01(0.15)	0.01(0.10)	-0.002(0.11)	----	----	0.70/0.86
PCK IRT	-0.11(0.11)	0.001(0.10)	-0.03(0.12)	0.12(0.11)	-0.01(0.12)	0.01(0.09)	0.02(0.15)	0.04(0.11)	0.02(0.11)	----	----	0.65/0.72

Notes: Sample size=187 teachers with complete data. Each outcome is measured with an overall scale (or percent correct for PCK) average as well as factor score (standardised). Standard errors in parentheses. All differences are measured in standard deviations (see Difference and Treatment effects columns). DD=simple test of significance for difference-in-difference (3b versus 3a). Model 1=regression with controls for treatment (T), endline (E) and treatment effect (T*E), Model 2=Model 1 with additional teacher characteristics (age, education, etc.). Bonf.=Bonferroni correction, Hoch=Hochberg correction. Attrition Null Hypothesis provides the p-values for IV-R and IV-P tests. See text and Section 3 for more details on methodology. *Difference between Control and Treated average is significantly different at $p \leq 0.05$; +Difference between Control and Treated average is significantly different at $p \leq 0.10$;

Table A3. Summary of grade 3 teacher questionnaire and PCK test results by sub-content area

Teacher outcome:	1. Baseline results:		2. Endline results:		3. Difference (SD) (2 minus 1)		4. Treatment effects:			Model 1 q-value correction:		Attrition null hypoth. (p-value)
	Cont. (99)	Treat (88)	Cont. (99)	Treat(88)	Cont.	Treat	DD	Model 1	Model 2	Bonf.	Hoch.	
Teacher questionnaire:												
2. Maths content and pedagogy												
2a. Numbers	4.1(0.07)	4.0(0.07)	4.2(0.05)	4.2(0.06)	-0.10(0.09)	0.09(0.10)	0.18(0.14)	0.19(0.14)	0.20(0.14)	1.00	0.91	0.38/0.57
2b. Measurement	4.1(0.08)	3.9*(0.08)	4.1(0.07)	4.1(0.07)	-0.19(0.11)	0.17*(0.10)	0.36*(0.14)	0.37*(0.15)	0.38*(0.15)	0.21	0.20	0.12/0.13
2c. Geometry	3.7(0.09)	3.6(0.10)	3.7(0.07)	3.9(0.08)	-0.15(0.08)	0.13*(0.10)	0.28*(0.13)	0.28*(0.14)	0.30*(0.14)	0.63	0.55	0.73/0.80
2d. Training Application	3.6(0.07)	3.0*(0.12)	3.7(0.06)	3.8(0.06)	-0.37(0.09)	0.33*(0.11)	0.71*(0.14)	0.91*(0.17)	0.94*(0.16)	0.01*	0.01*	0.01*/0.01*
3. Classroom Management												
3a. Positive Discipline	3.5(0.05)	3.3*(0.06)	3.4(0.06)	3.3(0.06)	-0.12(0.13)	0.10(0.13)	0.22(0.18)	0.24(0.20)	0.27(0.20)	----	----	0.05*/0.11
3b. Differentiation	3.1(0.06)	2.9*(0.05)	3.1(0.05)	3.0(0.05)	-0.13(0.14)	0.11(0.10)	0.24(0.17)	0.29(0.19)	0.30(0.19)	----	----	0.01*/0.02*
3c. Flexible grouping	3.3(0.03)	3.2(0.05)	3.3(0.05)	3.3(0.05)	-0.08(0.08)	0.08(0.11)	0.16(0.14)	0.20(0.18)	0.22(0.17)	----	----	0.11/0.01*
4. Knowledge, attitudes and practices												
4a. Self-belief	3.2(0.04)	3.1(0.05)	3.2(0.06)	3.2(0.04)	-0.12(0.10)	0.11(0.10)	0.23(0.14)	0.23(0.14)	0.27+(0.15)	1.00	0.91	0.01*/0.01*
4b. Attributes	3.9(0.03)	3.9(0.04)	3.9(0.03)	3.9(0.04)	0.04(0.013)	-0.04(0.12)	-0.08(0.17)	-0.10(0.19)	-0.13(0.19)	----	----	0.81/0.90
4c. Maths Anxiety	3.4(0.04)	3.3(0.04)	3.3(0.03)	3.3(0.03)	-0.05(0.11)	0.05(0.12)	0.10(0.17)	0.11(0.20)	0.07(0.22)	----	----	0.01*/0.02*

Table A3. Summary of grade 3 teacher questionnaire and PCK test results by sub-content area (continued)

Teacher outcome:	1. Baseline results:		2. Endline results:		3. Difference (SD) (2 minus 1)		4. Treatment effects:			Model 1 q-value correction:		Attrition null hypoth. (p-value)
	Cont. (99)	Treat (88)	Cont. (99)	Treat(88)	Cont.	Treat	DD	Model 1	Model 2	Bonf.	Hoch.	
6. MPDS												
6a. Affect	3.2(0.04)	3.1(0.05)	3.1(0.04)	3.1(0.04)	-0.09(0.13)	0.08(0.13)	0.18(0.18)	0.20(0.20)	0.15(0.20)	----	----	0.34/0.14
6e. Risk-taking	3.1(0.07)	3.1(0.04)	3.1(0.03)	3.1(0.05)	-0.10(0.14)	0.09(0.11)	0.19(0.18)	0.20(0.19)	0.15(0.19)	----	----	0.51/0.14
6f. Goals	3.3(0.04)	3.2(0.04)	3.3(0.05)	3.3(0.04)	-0.07(0.12)	0.06(0.11)	0.12(0.17)	0.14(0.20)	0.10(0.19)	----	----	0.83/0.65
6g. Motivation	3.4(0.06)	3.4(0.05)	3.3(0.05)	3.3(0.05)	0.01(0.11)	-0.01(0.12)	-0.02(0.16)	-0.02(0.21)	-0.02(0.21)	----	----	0.06+/0.01*
6h. Self-efficacy	2.9(0.04)	2.9(0.05)	3.0(0.03)	3.0(0.03)	-0.05(0.11)	0.04(0.09)	0.09(0.14)	0.09(0.14)	0.01(0.14)	----	----	0.90/0.18
Teacher PCK Test:												
Operations	64.3(2.7)	70.1(2.5)	65.4(3.3)	71.3(2.5)	-0.01(0.12)	0.01(0.10)	0.01(0.15)	0.01(0.13)	-0.02(0.14)	----	----	0.31/0.36
Fractions and number patterns	60.0(3.4)	59.6(2.4)	58.5(3.6)	58.8(3.3)	-0.02(0.10)	0.01(0.10)	0.03(0.15)	0.03(0.15)	0.02(0.16)	----	----	0.34/0.48
Problem solving	79.6(3.7)	76.8(3.5)	81.3(2.5)	81.3(3.1)	-0.04(0.11)	0.04(0.13)	0.08(0.17)	0.09(0.19)	0.09(0.20)	----	----	0.27/0.27
Measurement	59.1(2.6)	61.3(2.2)	62.2(1.8)	64.1(2.6)	0.01(0.10)	-0.01(0.09)	-0.02(0.13)	-0.01(0.14)	-0.02(0.16)	----	----	0.53/0.03*

Notes: Sample size=174 teachers with complete data. Each outcome is measured with an overall scale (or percent correct for PCK) average as well as factor score (standardised). Standard errors in parentheses. All differences are measured in standard deviations (see Difference and Treatment effects columns). DD=simple test of significance for difference-in-difference (3b versus 3a). Model 1=regression with controls for treatment (T), endline (E) and treatment effect (T*E), Model 2=Model 1 with additional teacher characteristics (age, education, etc.). Bonf.=Bonferroni correction, Hoch=Hochberg correction. Attrition Null Hypothesis provides the p-values for IV-R and IV-P tests. See text and Section 3 for more details on methodology.

+Difference between Control and Treated average is significantly different at $p \leq 0.10$;

*Difference between Control and Treated average is significantly different at $p \leq 0.05$

Table A4. Covariates of overall teacher questionnaire and PCK test averages

Teacher characteristic	Survey:	
	Teacher survey	PCK test
Teacher has complete data (base+end)	-0.04(0.09)	0.10(0.08)
Teacher is female	-0.26*(0.08)	-0.14+(0.08)
Teacher years of basic education	0.06+(0.03)	0.05*(0.02)
Teacher post basic education (reference=RTTC):		
Did not attend post basic ed	0.06(0.13)	-0.06(0.12)
Attended TVET	0.23(0.25)	0.33(0.26)
Attended PTTC	0.02(0.20)	-0.21(0.17)
Attended university	0.07(0.09)	0.19*(0.07)
Teacher experience (total)	0.01*(0.005)	-0.03*(0.004)
Tutors students after school	0.12(0.09)	0.05(0.06)
Teacher is contract teacher	-0.32*(0.10)	-0.33*(0.11)
Multigrade teacher	0.08(0.11)	0.12(0.10)
Treatment effects:		
Treatment province	-0.22*(0.10)	0.16*(0.08)
Endline survey	-0.06(0.06)	0.13*(0.06)
Treatment effect (treatment*endline)	0.24*(0.08)	0.02(0.07)
Teacher grade (ref=G1):		
Grade 2	0.03(0.09)	0.04(0.07)
Grade 3	0.21*(0.08)	-0.01(0.08)
Teacher questionnaire overall average	----	0.11*(0.03)
Teacher PCK overall average	0.12*(0.04)	----
Explained variance (R2)	0.09	0.19
Sample size	1,209	1,209

Data source: VVOB grade 1, 2 and 3 teacher surveys (2023–2025), baseline and endline

Notes: Sample averages provided (with standard deviations in parentheses). Dependent variables measure overall averages for teacher questionnaire (scale) and PCK test (percentage correct), and are standardised (z-scores). Coefficients obtained from Ordinary Least Squares (OLS) linear regression with standard errors (in parentheses) that are corrected for clustering at school level.

*Coefficient is statistically significant at $p \leq 0.05$ level;

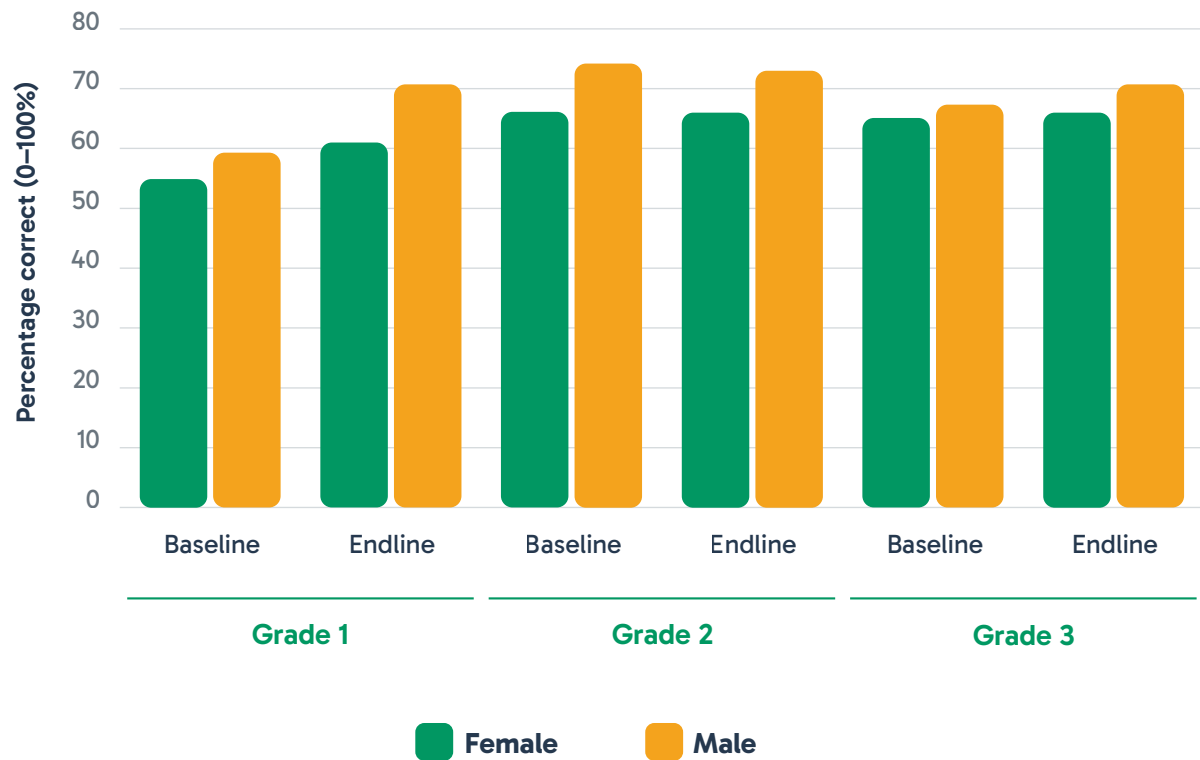
+Coefficient is statistically significant at $p \leq 0.10$ level;

Figure A2. Comparisons of baseline and endline overall teacher survey scores by gender, treatment schools only



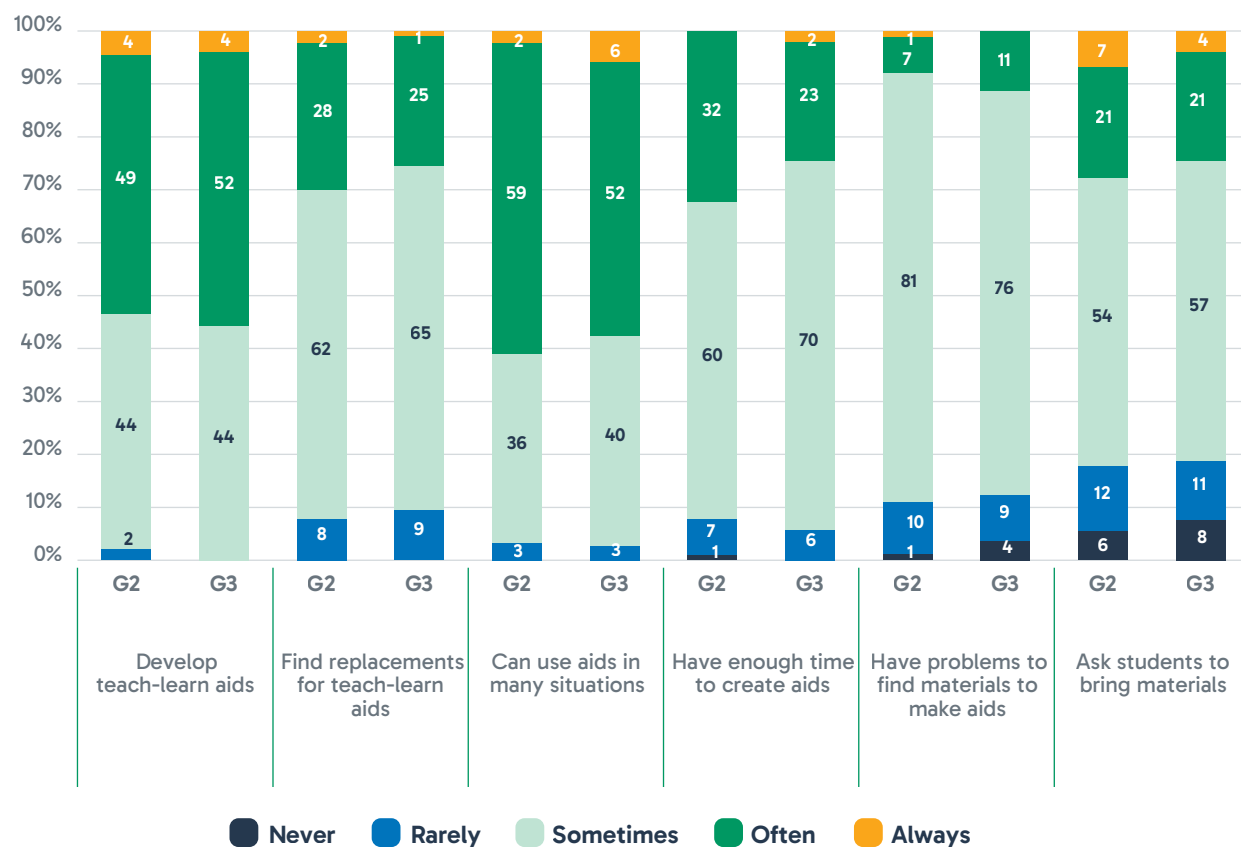
Data source: VVOB teacher survey, various years

Figure A3. Comparisons of baseline and endline teacher PCK percentage correct by gender, treatment schools only



Data source: VVOB teacher survey, various years

Figure A4. Grade 2 and 3 treatment province teacher responses about VVOB EGM programme participation experiences (“How often do you do these things?”)



Data source: VVOB teacher endline surveys, 2024–2025

Table A5. Summary of student EGMA test results for students with complete grade 1 baseline and grade 3 endline data

Test outcome:	1. Percent correct (baseline)		2. Percent correct (endline)		3. Standardised difference (z-score)		4. Treatment effects			Model 1 q-value correction		Attrition test p-values (IR/IP)
	Control	Treated	Control	Treated	Control	Treated	DD	Model 1	Model 2	Bonf.	Hoch.	
Number identification	20.7(1.8)	24.1(2.1)	93.7(1.0)	93.9(1.1)	0.06(0.06)	-0.06(0.08)	-0.12(0.11)	-0.09(0.08)	-0.08(0.08)	1.0	0.55	0.34/0.35
Number discrimination	15.3(1.6)	14.8(1.4)	92.8(1.1)	93.7(1.1)	-0.03(0.07)	0.03(0.06)	0.05(0.09)	0.04(0.06)	0.04(0.06)	1.0	0.55	0.81/0.40
Addition 1	5.2(1.0)	5.7(0.9)	55.6(1.5)	64.6*(2.4)	-0.18(0.06)	0.18*(0.09)	0.35*(0.11)	0.29*(0.09)	0.29*(0.09)	0.02*	0.02*	0.49/0.60
Subtraction 1	3.2(0.8)	1.7(0.3)	42.6(1.6)	50.9*(2.4)	-0.22(0.07)	0.22*(0.10)	0.43*(0.12)	0.38*(0.11)	0.39*(0.11)	0.01*	0.01*	0.14/0.14
Number pattern	34.5(1.8)	42.3*(2.7)	48.0(1.8)	56.0*(2.8)	-0.02(0.08)	0.02(0.09)	0.03(0.12)	0.01(0.09)	0.003(0.09)	0.18	0.13	----
Addition 2	17.7(1.8)	24.9*(2.6)	33.8(2.1)	42.5*(3.1)	-0.03(0.07)	0.03(0.07)	0.06(0.10)	0.05(0.09)	0.04(0.09)	0.23	0.14	----
Subtraction 2	6.2(0.8)	11.9*(2.2)	18.4(1.7)	24.7+(3.2)	-0.02(0.06)	0.02(0.09)	0.03(0.11)	0.02(0.10)	0.01(0.10)	0.91	0.33	----
Overall percent correct	10.4(1.1)	11.1(1.0)	61.2(1.2)	66.9*(1.8)	-0.15(0.07)	0.15*(0.09)	0.29*(0.11)	0.19*(0.07)	0.19*(0.07)	0.11	0.09+	0.54/0.64
Overall IRT score	-0.04(0.07)	0.09(0.07)	-0.16(0.06)	0.16*(0.10)	-0.10(0.07)	0.09(0.09)	0.19(0.12)	0.20(0.12)	0.21+(0.12)	1.0	0.33	0.26/0.04*
Global IRT score	-1.01(0.05)	-0.93(0.05)	0.68(0.04)	0.90*(0.07)	-0.11(0.07)	0.11+(0.09)	0.23+(0.12)	0.16+(0.08)	0.17*(0.08)	0.56	0.28	0.29/0.05*
Sample size	397	402	397	402	397	402	799	1,598	1,586	----	----	----

Data source: VVOB baseline and endline surveys, 2023–2025

Notes: Each EGMA test outcome is measured with an overall number of correct answers as well as a standardised (z-score). Standard errors (in parentheses) corrected for clustering at school level. All differences are measured in standard deviations (see Difference and Treatment effects columns). DD=difference-in-difference, Bonf.=Bonferroni correction, Hoch=Hochberg correction. See text and Section 3 for more details on methodology.

*Difference between Control and Treated average is significantly different at $p \leq 0.05$

+Difference between Control and Treated average is significantly different at $p \leq 0.10$

Table A6. Fixed effects estimations of VVOB programme impact by test area

Test outcome	G1 base + G3 end only		Data in all four moments	
	Female	Male	Female	Male
Number identification	-0.08(0.10)	-0.07(0.09)	-0.04(0.08)	-0.04(0.09)
Number discrimination	0.02(0.07)	0.07(0.07)	0.14*(0.06)	0.12(0.07)
Addition 1	0.28*(0.10)	0.31*(0.11)	0.17+(0.09)	0.24*(0.09)
Subtraction 1	0.35*(0.12)	0.44*(0.12)	0.18*(0.11)	0.34*(0.11)
Overall EGMA score (pct)	0.18*(0.08)	0.22*(0.09)	0.16*(0.08)	0.18*(0.08)
Overall EGMA score (IRT)	0.20(0.14)	0.24+(0.14)	0.14(0.11)	0.16(0.12)
Overall EGMA score (IRT global)	0.16(0.10)	0.19*(0.10)	0.10(0.08)	0.11(0.08)
Sample size	762	826	1,436	1,560

Data source: VVOB baseline and endline surveys, 2022–2024

Notes: Standard errors (in parentheses) corrected for clustering of students in schools.

*Parameter is statistically significant at $p \leq 0.05$

+Parameter is statistically significant at $p \leq 0.10$

