# Scientific Reasoning in Students with and Without Prior Exposure to Experiential Learning: A School-Adjusted Comparison

**Dr. Karuna K Simha[1*], Dr. Krithika M C[1]**

**[1]Prayoga Institute of Education Research, Bengaluru, Karnataka, India**

***Corresponding author**

**DOI: https://doi.org/10.47772/IJRISS.2026.100900709**

**Received: 27 September 2026; Accepted: 21 October 2026; Published: 21 October 2026**

## ABSTRACT

Experiential, hands-on science learning is now national policy in India, yet evidence on how such programmes relate to students' reasoning is thin. This study examined Kriya, a longitudinal experiential science programme that targets Piagetian reasoning operations in schools across Karnataka. We compared Grade 8 students who had taken part in Kriya during Grades 6 and 7 (continuing students, n = 2,202) with newly admitted classmates who had no prior exposure (n = 1,208) in 60 secondary schools across 34 districts. Both groups took the same 22-item test of reasoning operations set in science contexts in June 2026, at the start of Grade 8. The design is a non-equivalent comparison-group design without a pre-test, so all estimates describe associations. Continuing students scored 3.3 percentage points higher than new admissions (40.9% vs 37.6%; d = 0.18). Within schools, the difference fell to 2.1 points (0.11 SD). With standard errors clustered by school, its 95% confidence interval was 0.3 to 4.0 points. Competency-level differences rested on subscales of three or four items with low reliability, and none remained significant after Holm adjustment, so we treat them as exploratory. Continuing students were absent about three times as often as new admissions. Under a pessimistic assumption about absent students, the school-adjusted difference fell to 0.7 points and could not be distinguished from zero. No item showed meaningful differential item functioning. Even so, prior experience of the programme's assessments and differences between feeder schools remain the two principal alternative explanations, and the design cannot rule out either. Prior exposure to Kriya was therefore associated with a small and uncertain advantage, and both groups remained far from mastery. The study shows how a raw advantage shrinks once school composition, clustering and attendance are taken into account.

**Keywords:** Experiential learning; Scientific reasoning; Piagetian operations; Programme evaluation; Non-equivalent comparison groups

## INTRODUCTION

India's National Education Policy 2020 asks schools to adopt experiential, hands-on learning "as standard pedagogy within each subject" and to move assessment towards competencies (Ministry of Human Resource Development [MHRD], 2020, paras. 4.6, 4.34). The National Curriculum Framework for School Education builds on this by grounding middle-stage science in the exploration of concrete experience (National Council of Educational Research and Training [NCERT], 2023). The need is clear. In the most recent national survey, Grade 9 students answered 40% of science items correctly on average (NCERT, 2025a), and students in Karnataka answered 37% (NCERT, 2025b). Many students reach secondary school without the reasoning that school science assumes.

Non-governmental organisations have responded with a range of activity-based science programmes. Evidence on how such programmes relate to students' reasoning is thin, especially in government schools. Most rigorous evaluations in Indian schools have focused on foundational literacy and numeracy (Banerjee et al., 2007; Muralidharan et al., 2019). Few have examined science, and fewer still have measured reasoning rather than curricular content.

The evaluation problem is partly practical. Programmes run by non-governmental organisations are usually offered to whole schools or whole grades, so randomised comparison groups are rare. Programme teams do, however, collect large volumes of routine assessment data. These data are often reported as simple comparisons between participants and non-participants. Such comparisons can overstate or understate programme effects, depending on where participants are enrolled and who is present on test day (Shadish et al., 2002). How much a raw programme advantage changes under more careful analysis is itself a question of interest for evaluation.

International evidence gives reasons for both optimism and caution. Inquiry and discovery approaches raise learning when they are guided (Alfieri et al., 2011; Furtak et al., 2012; Lazonder & Harmsen, 2016). Programmes that target reasoning directly, such as Cognitive Acceleration through Science Education (CASE), have reported large and lasting effects (Adey & Shayer, 1990; Shayer & Adey, 1993). Yet a large randomised trial of a CASE successor in England found no effect on science attainment (Hanley et al., 2016). Effects also shrink as studies grow larger and use independent measures (Cheung & Slavin, 2016; Kraft, 2020), and many fade after programmes end (Bailey et al., 2017, 2020). Whether prior participation in an experiential programme in Indian government schools is associated with a measurable difference in reasoning is therefore an open empirical question. This study addresses that question for Kriya, a longitudinal experiential science programme that runs from Grade 6 to Grade 10 in schools across Karnataka. We use a comparison that arises naturally when students enter secondary school at Grade 8. Kriya secondary schools receive students from higher primary schools that ran Kriya in Grades 6 and 7, and students from schools that did not. Both groups sit the same reasoning test at the start of the year, in the same Grade 8 classrooms. The comparison holds school, classroom and test conditions constant. It is not experimental, however, and it does not support causal claims. The two groups come from different feeder schools, they have no common pre-test, and only one of them has sat the programme's assessments before. We therefore describe differences associated with prior exposure and analyse them in ways that make the main threats to a causal reading explicit.

**We ask four questions**

**RQ1.** Do students with two years of prior Kriya exposure score higher on a test of reasoning operations at entry to Grade 8 than newly admitted peers, and how large is the difference within schools?

**RQ2.** Which reasoning competencies, and which parts of the score distribution, show the difference?

**RQ3.** Does the difference vary across school contexts and programme-entry cohorts?

**RQ4.** How robust is the estimate to the clustering of students within schools, to differential absence, to test familiarity and to the quality of the instrument?

RQ1 is the primary question. It is answered by a single estimate, the school-adjusted difference on the whole test. RQ2 and RQ3 are secondary and exploratory, and RQ4 examines the robustness of the primary estimate.

The study makes three contributions. First, it provides a large-sample estimate of the difference in Piagetian reasoning associated with prior exposure to an experiential science programme in Indian schools, a setting where such estimates are scarce. Second, it shows how a raw programme advantage changes once school composition, the clustering of students within schools and differential attendance are taken into account. Third, it examines whether competency-level and subgroup results can bear the weight usually placed on them in programme reports. These lessons apply to any evaluation that relies on routine monitoring data.

## LITERATURE REVIEW

### Reasoning Operations as Targets of Science Education

Piagetian theory describes a set of logical operations that children construct as they act on the world. These

include conservation, classification, seriation, transitivity and reversibility, and later the combinatorial and proportional reasoning of formal thought (Inhelder & Piaget, 1958, 1964; Piaget & Inhelder, 1969). These operations are not science content. They are the reasoning that science content presumes. A student who cannot order objects along a dimension, or who does not see that a quantity is unchanged by a change in shape, will struggle with density, food chains or chemical change, however well these are taught.

The instrument used in this study assesses seven such operations. Classification is grouping objects by shared attributes. Decentration, assessed alongside it, is attending to more than one attribute at a time rather than centring on one. Seriation is ordering objects along a dimension. Transitivity is inferring a relation between two items from their relations to a third: if A is heavier than B and B is heavier than C, then A is heavier than C. Conservation is recognising that a quantity stays the same despite changes in appearance. Identity is recognising that an object or substance remains the same when nothing has been added or taken away. Inductive logic is drawing a general conclusion from specific observations, and generalisation is extending a rule to new cases. The first five competencies are characteristic of concrete operational thought. Induction and generalisation across contexts draw increasingly on the capacities of formal thought (Inhelder & Piaget, 1958; Piaget & Inhelder, 1969).

Piagetian cognitive operations and scientific reasoning are related but not synonymous. In the developmental literature, scientific reasoning usually refers to the skills involved in inquiry: generating hypotheses, designing experiments that control variables, and evaluating evidence (Zimmerman, 2007). Piagetian operations such as classification, seriation and conservation are logical structures on which these skills draw. The instrument used here assesses the operations, presented in science contexts. It does not assess experimentation or the evaluation of evidence directly. In this paper, scientific reasoning is therefore used in this narrower sense, as performance on reasoning operations embedded in science content.

Attainment of these operations is far from universal in adolescence. British norms showed that many students had not reached formal operations by the end of lower secondary school (Shayer et al., 1976). A replication three decades later found lower attainment than before (Shayer & Ginsburg, 2009). Group tests, such as Lawson's (1978) classroom test of formal reasoning, made it feasible to assess these operations at scale. For evaluation, the implication is that reasoning operations are a meaningful outcome in their own right. They are also a demanding one, because they develop slowly and are not easily taught to a test.

**Can Instruction Accelerate Reasoning?**

The strongest claims for teaching reasoning come from CASE. Over two years, students in lower secondary school worked through structured activities built on five pillars: concrete preparation, cognitive conflict, social construction, metacognition and bridging (Adey & Shayer, 1994; Venville & Oliver, 2015). Students made gains on Piagetian tests (Adey & Shayer, 1990).

Three years after the intervention ended, CASE students outperformed controls on national examinations in science, mathematics and English, with effect sizes of about 0.7 (Shayer & Adey, 1993; see also Shayer, 1999). Later adaptations reported gains in a low-socioeconomic school in regional Australia (Oliver et al., 2012) and in an academically selective school (Venville & Oliver, 2015). Long-term evaluation of these programmes has also raised methodological questions about comparison groups and attrition (Adey, 2005).

These results have not always travelled. A randomised trial of Let's Think Secondary Science in 53 English schools found no effect on science attainment (effect size = −0.02). A quarter of teachers taught six or fewer of the 19 lessons (Hanley et al., 2016). The gap between early efficacy studies and later scale-up is typical. Effect sizes in education fall as samples grow, as measures become independent of the programme, and as designs move from quasi-experimental to randomised (Cheung & Slavin, 2016). Implementation quality is a recurring explanation (Durlak & DuPre, 2008; O'Donnell, 2008).

The broader literature on inquiry and discovery learning points in the same direction. Syntheses find positive effects of inquiry-based science teaching (Minner et al., 2010), with a mean effect of about 0.50 and larger effects when teachers guide the inquiry (Furtak et al., 2012). Unassisted discovery is less effective than

explicit instruction. Enhanced discovery, with scaffolds and feedback, is more effective than other methods (Alfieri et al., 2011). Guidance also improves learning outcomes in inquiry settings (Lazonder & Harmsen, 2016). In low- and middle-income countries, programmes that change classroom pedagogy show larger effects than most other types of education intervention (Conn, 2017). Hands-on experience, in short, is a starting point. Whether it builds transferable reasoning depends on how it is structured and supported.

**Experiential Learning in Indian Classrooms**

Kolb (1984) described experiential learning as a cycle. Concrete experience is followed by reflective observation, abstract conceptualisation and active experimentation. This language now appears in Indian policy (MHRD, 2020; NCERT, 2023). Its uptake in government schools, however, takes place against low and uneven learning levels. Students in Indian middle schools are often several grade levels behind the curriculum, and the spread within a single grade is wide (Muralidharan et al., 2019).

Remedial and technology-aided programmes can raise achievement quickly, but gains may fade. In one well-known pair of trials, effects of 0.28 SD for remedial teaching and 0.47 SD for computer-assisted learning fell to about 0.10 SD a year later (Banerjee et al., 2007). A programme that aims to build reasoning, rather than to remediate content, faces a harder test: its effects must persist and transfer.

**Evaluating Programmes Without Random Assignment**

Most school programmes run by non-governmental organisations are not randomised. Evaluators often compare participants with whichever non-participants are available. Such non-equivalent comparison group designs face well-known threats, including selection, differential attrition, and confounding of the programme with the settings in which it runs (Shadish et al., 2002). Recent programme evaluations have used matching to strengthen such comparisons (Corning et al., 2023). When student background data are unavailable, however, evaluators must rely on design features and sensitivity analyses.

Two threats are central to the present study. First, participants and non-participants are rarely spread evenly across schools. Because schools differ substantially in achievement (Hedges & Hedberg, 2007), raw comparisons can mix programme effects with school effects. Second, groups may differ in who is present on test day. Absence is linked to lower attainment (Gottfried, 2010), so differential absence can bias the observed difference. Bounding methods make this second threat visible by asking how large an effect would be under plausible assumptions about missing students (Lee, 2009).

A third issue concerns the outcome measure. Programme teams increasingly use their own monitoring assessments to estimate effects. Such instruments are built to give feedback to schools, not to support causal claims, so their construct coverage, scoring and reliability need to be examined as part of the evaluation (Shadish et al., 2002). Item format also matters. Students can perform differently on hands-on and paper-based versions of science tasks (Chi et al., 2021), so a written test may capture only part of what experiential learning develops.

Interpreting effect size adds a further layer. Conventional benchmarks (Cohen, 1988) were not designed for education interventions. Benchmarks derived from field trials with achievement outcomes suggest that effects below 0.05 SD are small, 0.05 to 0.20 SD medium, and effects of 0.20 SD or more are large (Kraft, 2020). Effects should also be judged against normal growth, policy-relevant gaps and comparable programmes (Hill et al., 2008).

**Expectations for the Present Study**

The literature suggests three expectations. If prior exposure to Kriya is associated with stronger reasoning, the difference is likely to be modest at scale. It may be larger for operations that the programme's activities practise directly. And it is likely to vary with school context and implementation. Because the design cannot establish causal effects, we examine these expectations descriptively and make the main threats to inference explicit.

### The Kriya Programme

Kriya is a longitudinal experiential learning programme for Grades 6 to 10. It is implemented free of cost by Prayoga Institute of Education Research, a non-profit organisation, under memoranda of understanding with Samagra Shikshana Karnataka (SSK) and the Department of State Educational Research and Training (DSERT), the academic wing of the Department of Public Instruction in Karnataka. The programme aims to develop the reasoning operations that underpin school science: conservation, seriation, classification, transitivity, reversibility, identity, inductive logic and generalisation. Programme schools are grouped into four geographical zones (A to D). Kriya was piloted from June 2022 in ten private schools in Bengaluru's urban and rural areas and was launched across Karnataka in 2023–24. It runs in selected government and private schools. Regular science teachers act as Kriya facilitators, supported by Prayoga's academic coordinators. In 2026, the programme worked with 62 secondary schools and 93 facilitators.

Facilitators are trained through Prayoga's Teacher Empowerment Programme workshops in four areas: the pedagogy of experiential learning, managing resources in a hands-on classroom, classroom management, and science content. The workshops also build practical skills, such as conducting demonstrations, guiding students through experiments and handling equipment, so that hands-on work becomes part of everyday science teaching. Kriya course books present each reasoning competency through tasks. Each activity follows four phases. An orientation phase poses an everyday puzzle. A conceptualisation phase has students carry out a hands-on procedure. An investigation phase asks them to explain, compare and extend what they observed.

A conclusion phase states the underlying principle. Appendix B summarises two sample activities, on friction and on a lemon voltaic cell. In the friction activity, for example, students roll a cylinder onto three surfaces, compare the distances travelled and then order the surfaces by friction, which exercises seriation. Students are assessed three times each year on the programme's reasoning assessments. This study uses the baseline administration at the start of Grade 8. For continuing students, it reflects their reasoning after two years of participation in Grades 6 and 7, together with any familiarity with the assessment format gained from earlier administrations. For newly admitted students, it measures reasoning before any programme contact.

## METHODOLOGY

### Research Design

The study uses a non-equivalent comparison-group design without a pre-test (Shadish et al., 2002). Kriya runs in higher primary schools for Grades 6 and 7 and in secondary schools for Grades 8 to 10. Kriya secondary schools therefore receive two groups of Grade 8 entrants. Continuing students are students from Kriya higher primary schools, who carry a Kriya identifier; programme records call them Kriya-continuing students. New admissions are students from schools that did not run Kriya. We use these two terms throughout. Both groups were tested in June, at the start of the 2026–27 academic year. Because there is no common pre-test, the design estimates a difference in level, not in growth, and it does not support causal inference.

The comparison rests on an assumption: that without Kriya, continuing students would have scored like new admissions in the same secondary school. Some features support this assumption. Both groups enter the same Grade 8 classes, both have just moved from a higher primary school, and both come from schools serving the same local communities. Other features weaken it, as set out below. We therefore treat the estimates as the difference associated with prior exposure under this assumption. We use sensitivity analyses, and discuss the remaining threats, to show how conclusions change when parts of it are relaxed.

### Principal Threats to Internal Validity

Two alternative explanations could produce the observed difference without any contribution from the programme. We treat them as the principal threats to internal validity, not as minor limitations.

The first is selection of students and feeder schools. School fixed effects control for differences between receiving secondary schools, but they do not eliminate individual-level or feeder-school-level selection differences between continuing and newly admitted students. Within the same secondary school, the two groups may still differ in the quality of their previous school, socioeconomic background, earlier achievement, teacher quality, parental support, motivation, exposure to science and home language. The monitoring records contain no student background variables and no feeder-school identifier, so none of these can be adjusted for.

The second is test familiarity. Continuing students had sat the programme's reasoning assessments several times during Grades 6 and 7, whereas new admissions met the format for the first time. Retest gains on cognitive ability tests average about a third of a standard deviation (Scharfen et al., 2018). Because the school-adjusted difference is only about 0.11 SD and the instrument has just 22 items, even a modest familiarity effect could account for much or all of it. In this design, familiarity is confounded with programme exposure. We report the checks the data allow, differential item functioning and response completeness, but these cannot separate the two.

**Setting and Participants**

Kriya ran in 62 secondary schools in 2026–27. Sixty schools in 34 districts submitted Grade 8 records, listing 4,214 students on the roll. One school did not submit data because no academic coordinator was in post to administer the test and enter the results, and the other school was excluded. A total of 3,410 students (80.9% of the roll) attempted the test: 2,202 continuing students and 1,208 new admissions. These 3,410 students form the analytic sample for all score analyses. The attendance analyses use the 3,884 students still on the roll, that is, all 4,214 except the 330 recorded as transferred out. Table 1 shows how the roll reduced to the analytic sample. Four records were not scored. One student had been promoted to Grade 9 in the previous year and was therefore outside the Grade 8 cohort. Three students were recorded as having a disability or special educational needs. No responses were recorded for them, and the records do not show whether an adapted test was offered. Eighteen further records contained no response data and no recorded reason. Most test-takers attended government schools (86%) and rural schools (85%). New admissions formed a median of 27% of a school's Grade 8 cohort and outnumbered continuing students in 12 schools.

**Table 1. Disposition of the Grade 8 roll, by cohort**

| Status | Continuing students | New admissions | Total | % of roll |
|---|---|---|---|---|
| Attempted the test | 2,202 | 1,208 | 3,410 | 80.9 |
| Absent on test day | 371 | 66 | 437 | 10.4 |
| Transferred out | 330 | 0 | 330 | 7.8 |
| Long absent | 15 | 0 | 15 | 0.4 |
| No response data, no reason recorded | 14 | 4 | 18 | 0.4 |
| Disability or special needs recorded, no responses | 2 | 1 | 3 | 0.1 |
| Promoted to Grade 9 in the previous year | 1 | 0 | 1 | 0.0 |
| Total on roll | 2,935 | 1,279 | 4,214 | 100.0 |

Note. Continuing students carried a prior Kriya identifier from Grades 6 and 7. All score analyses use the 3,410 students who attempted the test. Attendance analyses use the 3,884 students not recorded as transferred out. Percentages may not sum to 100 because of rounding.

**Instrument**

The instrument comprised 22 multiple-choice items with four options each. Items covered physics (7 items), chemistry (8) and biology (7). Each item was mapped to one of seven reasoning competencies: classification and decentration, transitivity, seriation, conservation, inductive logic and identity (three items each), and generalisation (four items). Table 2 shows the test blueprint and how it maps onto the programme's stated competencies. Kriya targets eight competencies, and seven are assessed. Reversibility, although a programme aim, is not assessed in this form. Decentration is assessed together with classification rather than as a separate competency. Box 1 shows a sample conservation item.

**Table 2. Test blueprint and its relation to the programme's stated competencies**

| Competency | Kriya aim | Physics | Chemistry | Biology | Total | Item numbers |
|---|---|---|---|---|---|---|
| Classification and decentration | Yes | 1 | 1 | 1 | 3 | 1, 9, 16 |
| Transitivity | Yes | 1 | 1 | 1 | 3 | 4, 12, 19 |
| Seriation | Yes | 1 | 1 | 1 | 3 | 3, 11, 18 |
| Conservation | Yes | 1 | 1 | 1 | 3 | 6, 14, 21 |
| Inductive logic | Yes | 1 | 1 | 1 | 3 | 5, 13, 20 |
| Identity | Yes | 1 | 1 | 1 | 3 | 2, 10, 17 |
| Generalisation | Yes | 1 | 2 | 1 | 4 | 7, 8, 15, 22 |
| Reversibility | Yes | 0 | 0 | 0 | 0 | Not assessed |
| Total | | 7 | 8 | 7 | 22 | |

Note. Kriya aim indicates whether the competency is one of the eight the programme states it targets. Decentration is not a separately stated aim; the instrument assesses it with classification. Item-level statistics are given in Appendix A (Table A1).

**Box 1.** Sample item to test conservation competency

A student dissolves 10 g of salt in a beaker containing 100 mL of water. She then pours the solution into a China dish and heats it on a hot plate to evaporate the water. The amount of salt obtained after evaporation will be:

A. less than the amount initially dissolved in water

B. the same as the amount dissolved initially

C. less, as some salt changes to water and evaporates

D. more than the amount taken initially for dissolution

Correct answer: B. The item assesses conservation of the amount of substance through a change of state.

Internal consistency of the whole test was adequate for group comparisons (KR-20 = .72, 95% CI [.71, .74]; Kuder & Richardson, 1937; Feldt, 1965), but below the level usually sought for decisions about individual students. The standard error of measurement was 2.1 items. Item difficulty ranged from 23.3% to 54.4% correct, with no ceiling effects. Against the commonly used minimum of .20 for item discrimination (Ebel & Frisbie, 1991), three items, all in chemistry, fell far short, with corrected point-biserial correlations of .00, .07 and .08. A fourth item (identity, physics) was marginal at .17. The competency subscales were much less reliable than the whole test. Their KR-20 values ranged from .03 (identity) to .45 (seriation), and correlations between subscale scores ranged from .15 to .42. Each competency is therefore inferred from

three or four items, several of them weakly discriminating, and the subscales cannot be treated as separate, reliably measured constructs. We retained all items to preserve the instrument as administered and examine their influence in the Results. Item statistics are given in Appendix A.

Random guessing yields an expected score of 5.5 out of 22 (25%). We use this as a reference point. A student scoring at or below 5.5 has performed no better than guessing.

**Data Processing**

All responses were re-scored against the answer key rather than taken from pre-computed totals. This identified 84 records in which stray spaces in response cells had caused correct answers to be scored as wrong, understating totals by 1 to 5 marks. These were corrected. Unanswered items (1,086 responses, 1.4% of all responses), 107 responses with more than one option marked and 8 uninterpretable entries were scored as incorrect, following convention for single-best-answer tests. Records coded as transferred out, absent, long absent or without response data were excluded from score analyses but retained for the attendance analysis, as were the four records described above. School and district names were standardised before aggregation.

**Data Analysis**

For RQ1, we compared group means with Welch's t-test (Welch, 1947) and checked the result with the Mann–Whitney U test. We report Cohen's d with a 95% confidence interval and the probability of superiority, that is, the chance that a randomly chosen continuing student outscores a randomly chosen new admission (McGraw & Wong, 1992). To remove differences in school composition, we estimated the ordinary least squares model

$$Y_{is} = \alpha_s + \beta K_{is} + \varepsilon_{is} \quad (1)$$

where $Y_{is}$ is the percentage score of students i in school s, $\alpha_s$ is a school fixed effect, $K_{is}$ equals 1 for continuing students and 0 for new admissions, and $\varepsilon_{is}$ is an error term. The coefficient $\beta$ is the average within-school difference. To compare the importance of school and cohort, we report the proportion of score variance explained by school indicators alone and by the cohort indicator alone.

Students are nested in 60 schools, and within-school differences varied from school to school. We therefore report standard errors clustered by school (CR1, with confidence intervals and tests based on the t distribution with 59 degrees of freedom; Cameron & Miller, 2015). For comparison, we also report conventional and heteroscedasticity-robust standard errors, a random-intercept multilevel model, and a wild cluster bootstrap p value (9,999 replications, Rademacher weights, null imposed; Cameron et al., 2008), because 60 clusters is a moderate number. Unless stated otherwise, all confidence intervals for school-adjusted estimates use school-clustered standard errors. To convey practical meaning, we express the primary estimate in items, in standard deviation units and as Cohen's $U_3$, the share of new admissions whom the average continuing student would outscore.

For RQ2, we compared score distributions across four bands (0–25%, 25–50%, 50–75% and 75–100% correct) using a chi-square test, and estimated school-adjusted differences in the share of students in each band with linear probability versions of Equation 1. We estimated Equation 1 separately for each competency and subject strand, with the mean item score as the outcome. Because there were seven competency outcomes, we adjusted p values with the Holm (1979) procedure, which controls the familywise error rate, and report Benjamini–Hochberg (1995) false discovery rate values alongside. We also tested whether the seven school-adjusted competency differences were equal, using a stacked model with school-clustered standard errors. We treat the competency analyses as exploratory and emphasise effect sizes and confidence intervals rather than significance.

For RQ3, we computed within-school differences for the 54 schools that had both groups and used a two-sided sign test to compare the number of schools with positive and negative differences. We described

differences by school management, board, location and zone, reporting both raw and school-adjusted differences with school-clustered 95% confidence intervals, and tested whether the adjusted difference varied across zones with a cohort-by-zone interaction. These subgroup results are descriptive, because several groups contain few schools. We also grouped continuing students by the academic year encoded in their Kriya identifier (2022–23, 2023–24 or 2024–25), which the programme's records use to mark the year a student entered the Kriya cohort, and compared each group with new admissions.

These years follow the programme's rollout. The 2022–23 identifiers occur only in two private schools, consistent with the pilot; the 2023–24 identifiers belong to the statewide launch; and the 2024–25 identifiers occur in a single school. We therefore treat the identifier year as a marker of programme-entry cohort, not as a count of years of participation, and report these comparisons as exploratory descriptive differences, not as evidence about programme dosage.

For RQ4, we first compared absence rates among students still on the roll. We then imputed a score for each absent or long-absent student and re-estimated both the raw gap and the school-adjusted difference. Two scenarios were used. In the first, absent students were assigned the 25th percentile of their own cohort's observed distribution. This is a moderately pessimistic assumption: it treats absent students as resembling the weakest quarter of those who attended, in line with evidence that absence is associated with lower attainment (Gottfried, 2010).

In the second, they were assigned the 10th percentile, a stringent assumption that places them below the chance score of 5.5 items. Because any fixed percentile is arbitrary, we also report a tipping-point analysis, which finds the score absent continuing students would need for the adjusted difference to vanish or lose significance, and Lee (2009) trimming bounds for the raw gap. Lee bounds require no assumption about absent students' scores but assume that being a continuing student could only reduce, never increase, a student's chance of being tested.

To check for test familiarity, we ran a Mantel–Haenszel differential item functioning (DIF) analysis on all 22 items, matching students on total score, and classified items with the Educational Testing Service A/B/C rules (Holland & Thayer, 1988; Zwick, 2012). We also compared the share of unanswered and multiply marked responses, one observable sign of unfamiliarity with the format. Finally, we examined whether competency-level results coincided with poorly functioning items.

Percentages refer to items correct unless stated otherwise, and differences are in percentage points (pp). Standardised effects use the pooled standard deviation of total scores (4.05 items, or 18.4 pp).

### Ethical Approval

The Kriya main study was approved by the Ethics Committee of Prayoga Institute of Education Research (reference ER-23-11-09). Samagra Shikshana Karnataka and DSERT permitted the programme and its assessments on behalf of schools. This study is a secondary analysis of anonymised assessment records collected during routine programme monitoring.

## RESULTS

### Overall Difference (RQ1)

Continuing students scored higher than new admissions (Table 3). Their mean was 9.00 out of 22 (40.9%), against 8.27 (37.6%) for new admissions. The difference was 0.73 items, or 3.3 pp, 95% CI [2.0, 4.6]. It was statistically significant, $t(2493) = 5.04$, $p < .001$, and the Mann–Whitney test agreed ($p < .001$). The standardised difference was $d = 0.18$, 95% CI [0.11, 0.25], just below Cohen's (1988) threshold for a small effect. The probability of superiority was .56. In other words, a continuing student outscored a new admission slightly more than half the time.

**Table 3. Score distribution and group comparison at entry to Grade 8**

| Measure | Continuing students (n = 2,202) | New admissions (n = 1,208) | Raw difference [95% CI] | School-adjusted difference [95% CI] |
|---|---|---|---|---|
| Mean score (of 22) | 9.00 | 8.27 | 0.73 [0.45, 1.01] | 0.47 [0.06, 0.87] |
| Mean percentage correct | 40.9 | 37.6 | 3.3 [2.0, 4.6] | 2.1 [0.3, 4.0] |
| Standard deviation (items) | 4.06 | 4.04 | | |
| Median (items) | 8 | 7 | 1 | |
| 25th percentile (items) | 6 | 5 | 1 | |
| 75th percentile (items) | 12 | 11 | 1 | |
| 90th percentile (items) | 15 | 15 | 0 | |
| Share of cohort in each band (%) | | | | |
| 0–25% (at or below chance) | 21.8 | 27.3 | −5.6 [−8.6, −2.5] | −1.3 [−6.0, 3.3] |
| 25–50% | 51.2 | 52.6 | −1.5 [−5.0, 2.0] | −4.0 [−9.6, 1.7] |
| 50–75% | 22.6 | 14.9 | 7.7 [5.1, 10.4] | 6.0 [0.5, 11.4] |
| 75–100% | 4.5 | 5.1 | −0.7 [−2.2, 0.8] | −0.7 [−4.6, 3.3] |
| 14 or more items correct | 16.9 | 13.0 | 3.9 [1.5, 6.4] | 3.6 [0.4, 6.9] |
| Effect sizes | | | | |
| Cohen's d (pooled SD) | | | 0.18 [0.11, 0.25] | 0.11 [0.01, 0.21] |
| Probability of superiority | | | .56 | |

Note. Differences are in percentage points (pp) unless stated otherwise and were computed from unrounded values. Raw-difference intervals are Welch intervals. School-adjusted differences are cohort coefficients from ordinary least squares models with school fixed effects (Equation 1); for bands, the outcome is a 0/100 indicator. Their 95% CIs use standard errors clustered by school (60 schools; t with 59 df). Chance = 5.5 of 22 items. Welch $t(2493) = 5.04$, $p < .001$; Mann–Whitney $p < .001$.

Part of the raw difference reflected where the two groups were enrolled. School mean scores ranged from 21% to 72%. School membership explained 52.9% of the variance in individual scores, while cohort membership explained 0.7%. With school fixed effects, the difference fell to 2.11 pp. Roughly a third of the raw gap was therefore due to school composition. We treat the within-school difference as the preferred estimate.

Table 4 shows how the uncertainty around this estimate depends on how the clustering of students within schools is handled. With conventional standard errors, the 95% CI was [1.11, 3.12]. Clustering by school nearly doubled the standard error, from 0.51 to 0.92, because the within-school difference varied considerably between schools (standard deviation of school-level gaps = 7.0 pp). The clustered 95% CI was [0.27, 3.96], $t(59) = 2.29$, $p = .025$, and the wild cluster bootstrap gave $p = .027$.

A random-intercept model reproduced the point estimate but, like the conventional model, assumes a common within-school difference and so gives a narrower interval. In standardised terms, the clustered estimate corresponds to 0.11 SD, 95% CI [0.01, 0.21]. The difference therefore remained positive, but the lower end of its interval was close to zero. The estimate was similar when the single-school 2024–25 cohort was excluded (2.15 pp) and when the analysis was restricted to government schools (1.94 pp).

**Table 4. Sensitivity of the school-adjusted difference to the standard-error specification and the sample**

| Specification | Estimate (pp) | SE | 95% CI | p |
|---|---|---|---|---|
| Panel A. Analytic sample (N = 3,410; 60 schools) | | | | |
| School fixed effects, conventional SE | 2.11 | 0.51 | [1.11, 3.12] | < .001 |
| School fixed effects, heteroscedasticity-robust (HC1) SE | 2.11 | 0.50 | [1.14, 3.09] | < .001 |
| School fixed effects, school-clustered (CR1) SE | 2.11 | 0.92 | [0.27, 3.96] | .025 |
| School fixed effects, wild cluster bootstrap | 2.11 | | | .027 |
| Random-intercept multilevel model | 2.11 | 0.51 | [1.11, 3.12] | < .001 |
| Panel B. Alternative samples, school-clustered SE | | | | |
| Excluding the 2024–25 single-school cohort (N = 3,321) | 2.15 | 1.01 | [0.12, 4.18] | .039 |
| Government schools only (N = 2,948; 49 schools) | 1.94 | 0.69 | [0.55, 3.32] | .007 |

Note. Estimates are percentage points (pp). Clustered intervals and tests use the t distribution with G − 1 degrees of freedom, where G is the number of schools. The wild cluster bootstrap used 9,999 replications with Rademacher weights and the null hypothesis imposed. The random-intercept model was estimated by restricted maximum likelihood (intraclass correlation = .48).

In practical terms, 2.11 pp is 0.47 of an item on a 22-item test, or about one additional correct answer for every two students. If scores are roughly normal, the average continuing student would outscore about 55% of new admissions in the same school (Cohen's $U_3$ = .55), against 50% if there were no difference. The difference is also less than a quarter of the test's standard error of measurement (2.1 items), so it is detectable for groups of students but not for individuals. Field-trial benchmarks would place 0.11 SD in the medium range for education interventions (Kraft, 2020), but those benchmarks come from causal designs, and the clustered interval runs from negligible to about 0.2 SD.

**Where the Difference Lies (RQ2, Exploratory)**

The score distributions of the two groups overlapped substantially but differed in shape (Figure 1), $\chi^2(3) = 34.63$, $p < .001$, Cramér's V = .10. In raw terms, fewer continuing students scored at or below chance (21.8% vs 27.3%; difference = −5.6 pp, 95% CI [−8.6, −2.5]), and more scored in the 50–75% band (22.6% vs 14.9%; 7.7 pp [5.1, 10.4]). There was no difference in the top band (4.5% vs 5.1%), and both groups had the same 90th percentile (15 of 22). Within schools, the pattern changed (Table 3).

The difference at the chance floor shrank to −1.3 pp, 95% CI [−6.0, 3.3], so the raw floor difference largely reflected school composition. The shift into the 50–75% band remained (6.0 pp [0.5, 11.4]), as did the difference in the share scoring 14 or more items (3.6 pp [0.4, 6.9]). We present this as an observed distributional pattern. It does not show that prior exposure particularly benefits weaker or stronger students. Differential attendance, test familiarity, the floor and ceiling properties of a short test and measurement error could each shape where in the distribution a difference appears.

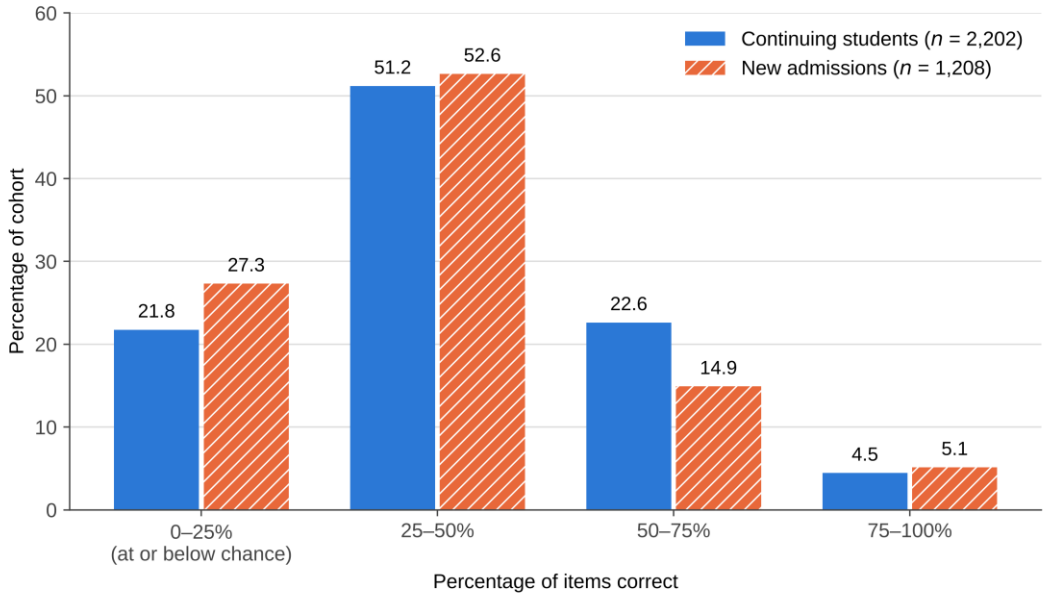



**Figure 1.** Distribution of students across performance bands, by cohort. Bands show the percentage of the 22 items answered correctly; the lowest band corresponds to scores at or below chance (5.5 of 22). Values are percentages of each cohort (raw, not school-adjusted).

At the competency level, school-adjusted differences ranged from −0.9 pp (identity) to +3.8 pp (classification and decentration) (Table 5; Figure 2). With school-clustered standard errors, the intervals for classification and decentration (+3.8 pp, 95% CI [0.7, 6.9]) and conservation (+2.5 pp [0.1, 5.0]) excluded zero. Neither difference remained significant after Holm adjustment for seven outcomes (adjusted p = .12 and .26), and none of the seven reached significance at a false discovery rate of 10%. The intervals for transitivity (+3.4 pp), seriation (+2.9 pp), inductive logic (+2.5 pp), generalisation (+0.9 pp) and identity (−0.9 pp) all included zero. A joint test indicated that the seven differences were not all equal, F(6, 59) = 2.44, p = .036. This heterogeneity came from identity: without it, the remaining six differences were consistent with a common value, F(5, 59) = 0.82, p = .54. With subscale reliabilities between .03 and .45, a single weak item can shift a competency score substantially. We therefore treat the competency profile as exploratory and do not rank the competencies.

**Table 5. Performance by reasoning competency and subject strand**

| | Items | Continuing (%) | New admissions (%) | Raw gap (pp) | School-adjusted gap (pp) [95% CI] | p | Holm p | KR-20 |
|---|---|---|---|---|---|---|---|---|
| Panel A. Reasoning competency (exploratory) | | | | | | | | |
| Classification and decentration | 3 | 34.2 | 29.5 | 4.7 | 3.8 [0.7, 6.9] | 0.018 | 0.12 | 0.11 |
| Transitivity | 3 | 43.9 | 38.8 | 5.1 | 3.4 [−0.4, 7.3] | 0.078 | 0.39 | 0.32 |
| Seriation | 3 | 51.5 | 46.5 | 5 | 2.9 [−1.0, 6.8] | 0.15 | 0.59 | 0.45 |
| Conservation | 3 | 36.6 | 33.7 | 2.9 | 2.5 [0.1, 5.0] | 0.043 | 0.26 | 0.22 |
| Inductive logic | 3 | 47.4 | 42.8 | 4.6 | 2.5 [−1.0, 6.1] | 0.16 | 0.59 | 0.33 |
| Generalisation | 4 | 34.4 | 32.8 | 1.6 | 0.9 [−0.8, 2.6] | 0.28 | 0.59 | 0.34 |
| Identity | 3 | 40.6 | 40.7 | −0.1 | −0.9 [−3.7, 1.9] | 0.52 | 0.59 | 0.03 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| Panel B. Subject strand (descriptive) | | | | | | | | |
| Physics | 7 | 44.6 | 41.4 | 3.2 | 2.6 [−0.2, 5.3] | 0.069 | | 0.48 |
| Biology | 7 | 41.6 | 36.7 | 4.9 | 3.2 [0.6, 5.8] | 0.018 | | 0.64 |
| Chemistry | 8 | 37.1 | 35.1 | 2 | 0.8 [−1.5, 3.1] | 0.49 | | 0.37 |

Note. Adjusted gaps are from separate ordinary least squares models with school fixed effects, using each competency's or strand's mean item score as the outcome. Intervals and p values use school-clustered standard errors (t, 59 df). Holm p adjusts across the seven competencies; Benjamini–Hochberg adjusted values ranged from .12 to .52. KR-20 was computed in the analytic sample. Test that the seven competency gaps are equal: F(6, 59) = 2.44, p = .036; excluding identity: F(5, 59) = 0.82, p = .54. pp = percentage points.

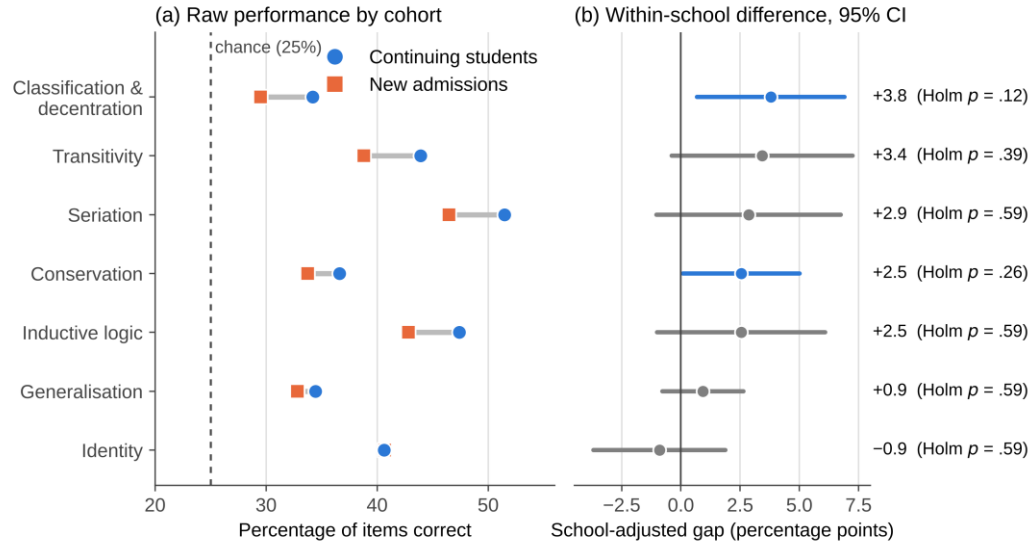



**Figure 2. Results by reasoning competency.**

(a) Percentage of items correct, by cohort; the dashed line marks chance. (b) Within-school difference from school fixed-effects models with school-clustered 95% confidence intervals; blue intervals exclude zero before adjustment for multiple comparisons, grey intervals include zero. Labels give Holm-adjusted p values.

Absolute levels were low for both groups. The highest competency mean was 51.5% (continuing students, seriation). Classification and decentration and generalisation were the hardest competencies for both groups, with means at or below 34.4%.

By subject strand, the raw difference was largest in biology (+4.9 pp), followed by physics (+3.2 pp) and chemistry (+2.0 pp). Within schools, the corresponding differences were 3.2, 2.6 and 0.8 pp, and only the biology interval excluded zero. Chemistry was the weakest strand for both groups.

**Variation across Schools, Contexts and Programme-Entry Cohorts (RQ3, Exploratory)**

Of the 54 schools with both groups, continuing students scored higher in 34 (63%). This majority was not significantly different from an even split (two-sided sign test, p = .076). Among the 34 schools with at least ten students in each group, continuing students led in 22 (p = .12). Ten of these schools showed an advantage above 5 pp for continuing students, and five showed an advantage above 2 pp for new admissions. The widest contrast involved two schools in the same network. One showed a 13.6 pp advantage for continuing students; the other showed a 15.9 pp advantage for new admissions. Across all 54 schools, the standard deviation of within-school gaps was 7.0 pp.

Table 6 reports raw and school-adjusted differences by context, with school-clustered intervals. These results are descriptive. Several groups contain few schools, and raw gaps within a group still mix school composition with the cohort difference. Raw gaps were larger in private (8.7 pp) than in government schools (2.5 pp). Within schools, however, the differences were similar (3.3 and 1.9 pp; test of the difference, p = .81), and the private-school interval was very wide [−9.4, 16.1] because only nine private schools had both groups.

The two CBSE schools are too few for any interval to be informative. Rural and urban schools showed similar adjusted differences (2.3 and 1.2 pp). By zone, the raw gap was largest in Zones A (+6.4 pp) and B (+5.1 pp) and reversed in Zone D (−2.5 pp). Within schools, the adjusted differences for Zones A to D were 2.9, 3.8, 0.8 and 1.0 pp. All four were positive, and they did not differ significantly from one another, F(3, 59) = 0.65, p = .59. The raw reversal in Zone D therefore reflected where the two groups were enrolled, not a within-school disadvantage. Among the 20 districts with at least 20 students in each group, raw gaps ranged from +14.5 to −5.4 pp.

**Table 6. Variation in the cohort difference by school context, zone and programme-entry cohort (descriptive)**

| | **Schools (with both groups)** | **Continuing n (mean %)** | **New admissions n (mean %)** | **Raw gap (pp) [95% CI]** | **School-adjusted gap (pp) [95% CI]** |
|---|---|---|---|---|---|
| Panel A. School context | | | | | |
| Government | 49 (45) | 1,908 (39.0) | 1,040 (36.4) | 2.5 [−3.0, 8.1] | 1.9 [0.5, 3.3] |
| Private | 11 (9) | 294 (53.6) | 168 (44.9) | 8.7 [0.9, 16.5] | 3.3 [−9.4, 16.1] |
| State board | 58 (52) | 2,135 (40.6) | 1,149 (37.3) | 3.3 [−2.0, 8.7] | 1.7 [0.0, 3.4] |
| CBSE | 2 (2) | 67 (51.8) | 59 (44.5) | 7.4 [n.e.] | n.e. |
| Rural | 50 (45) | 1,859 (41.0) | 1,026 (37.9) | 3.1 [−2.7, 9.0] | 2.3 [0.2, 4.4] |
| Urban | 10 (9) | 343 (40.6) | 182 (36.1) | 4.5 [−6.5, 15.6] | 1.2 [−2.3, 4.8] |
| Panel B. Zone | | | | | |
| Zone A | 21 (19) | 893 (38.0) | 455 (31.6) | 6.4 [−1.0, 13.8] | 2.9 [1.0, 4.7] |
| Zone B | 6 (6) | 261 (35.6) | 188 (30.5) | 5.1 [0.0, 10.2] | 3.8 [−0.9, 8.6] |
| Zone C | 17 (14) | 573 (42.2) | 293 (39.8) | 2.4 [−3.2, 8.0] | 0.8 [−5.9, 7.4] |
| Zone D | 16 (15) | 475 (47.8) | 272 (50.2) | −2.5 [−12.9, 8.0] | 1.0 [−2.4, 4.5] |
| Panel C. Programme-entry cohort (year encoded in identifier) | | | | | |
| 2022–23 pilot cohort | 2 | 29 (47.3) | 1,208 (37.6) | 9.7 [3.6, | n.e. |

| | | | | | |
|---|---|---|---|---|---|
| (two private schools) | | | | 15.9] | |
| 2023–24 statewide launch cohort | 57 (51) | 2,084 (41.6) | 1,208 (37.6) | 4.0 [2.7, 5.3] | 1.8 [−0.1, 3.8] |
| 2024–25 cohort (one school) | 1 | 89 (22.0) | 1,208 (37.6) | −15.6 [−17.8, −13.4] | 1.8 [−0.5, 4.1] |

Note. Raw gaps in Panels A and B use school-clustered intervals; school-adjusted gaps come from school fixed-effects models with school-clustered intervals and use only schools with both groups. In Panel C, raw gaps compare each cohort with all new admissions (Welch intervals).

The adjusted gap for 2023–24 compares the cohort with new admissions in the same schools. For 2024–25, it compares the 89 students with the 154 new admissions in the same school (heteroscedasticity-robust interval). The two pilot schools enrolled only 11 new admissions. n.e. = not estimable with an informative interval (two schools or fewer); CBSE = Central Board of Secondary Education; pp = percentage points. Gaps were computed from unrounded means. Panel C reports exploratory descriptive differences by programme-entry cohort. They should not be read as evidence about programme dosage.

The 2022–23 pilot cohort (n = 29) came from two private schools that enrolled only 11 new admissions, so its raw lead of 9.7 pp cannot be separated from school type. The 2024–25 cohort (n = 89) came from a single school, and its raw gap of −15.6 pp reflected that school's low overall performance. Within the school, its continuing students scored 1.8 pp above its 154 new admissions.

The 2023–24 launch cohort contains about 95% of continuing students, and its school-adjusted difference (1.8 pp, 95% CI [−0.1, 3.8]) was close to the overall estimate. The cohort data therefore do not support any claim that longer or shorter participation is associated with larger or smaller differences.

**Attendance and Robustness (RQ4)**

Attendance differed sharply between the groups (Table 7, Panel A). Of 2,935 continuing students on the roll, 330 (11.2%) had transferred out. Among those still on the roll, 14.8% were absent on test day or long absent, against 5.2% of new admissions, odds ratio = 3.20, 95% CI [2.44, 4.19], $\chi^2(1) = 77.8$, $p < .001$. Transfers were concentrated. Four schools accounted for 136 of the 330, and in these schools, transfers made up between 25% and 54% of the Grade 8 roll.

If absent students would have scored below those who attended, the observed gap overstates the difference among all students on the roll. Table 7 (Panel B) shows the effect of imputing scores for absent students. At their cohort's 25th percentile (6 items for continuing students and 5 for new admissions), the raw gap fell from 3.3 to 2.1 pp, and the school-adjusted difference fell from 2.11 to 1.71 pp, 95% CI [0.12, 3.31]. At the 10th percentile (4 items for both cohorts, below the chance score of 5.5), the raw gap fell to 0.9 pp and the adjusted difference to 0.65 pp, 95% CI [−1.03, 2.33], $p = .44$.

Because both cohorts receive the same imputed score in this scenario, it does not carry the observed advantage into the imputed values. The tipping-point analysis gives a fuller picture. Holding absent new admissions at their own 25th percentile, the adjusted difference stays significantly above zero only if absent continuing students would have scored at least about 6 of 22 items, roughly the lower quartile of continuing students who sat the test.

It reaches zero if they would have scored about 3.3 items. Lee (2009) bounds on the raw gap, which trim 10.3% of new admissions to equalise the share tested, ranged from 0.5 to 7.6 pp. Figure 3 summarises the estimate across specifications. The point estimate stayed positive in every scenario, but under the most pessimistic assumption it could not be distinguished from zero.

**Table 7. Attendance by cohort and sensitivity of the cohort difference to differential absence**

| | Continuing students | New admissions | Raw gap (pp) | Raw gap (SD) | School-adjusted gap (pp) [95% CI] |
|---|---|---|---|---|---|
| Panel A. Attendance | | | | | |
| On roll | 2,935 | 1,279 | | | |
| Transferred out, n (%) | 330 (11.2) | 0 (0.0) | | | |
| Absent or long absent, n | 386 | 66 | | | |
| Absence rate among students still on roll (%) | 14.8 | 5.2 | | | |
| Tested | 2,202 | 1,208 | | | |
| Panel B. Mean percentage correct under alternative assumptions | | | | | |
| Observed (absent students excluded) | 40.9 | 37.6 | 3.3 | 0.18 | 2.11 [0.27, 3.96] |
| Absent students at own cohort's 25th percentile | 38.9 | 36.8 | 2.1 | 0.11 | 1.71 [0.12, 3.31] |
| Absent students at own cohort's 10th percentile | 37.5 | 36.6 | 0.9 | 0.05 | 0.65 [−1.03, 2.33] |
| Lee (2009) bounds, raw gap | | | [0.5, 7.6] | [0.02, 0.42] | |

Note. Absence among students still on roll: odds ratio = 3.20, 95% CI [2.44, 4.19], $\chi^2(1) = 77.8$, $p < .001$. Imputed scores were 6 (continuing) and 5 (new admissions) items at the 25th percentile and 4 items for both cohorts at the 10th percentile. School-adjusted gaps use school-clustered intervals. Lee bounds trim 10.3% of tested new admissions from the top or bottom of their distribution. Standardised gaps use the pooled standard deviation of total scores (18.4 pp). pp = percentage points.

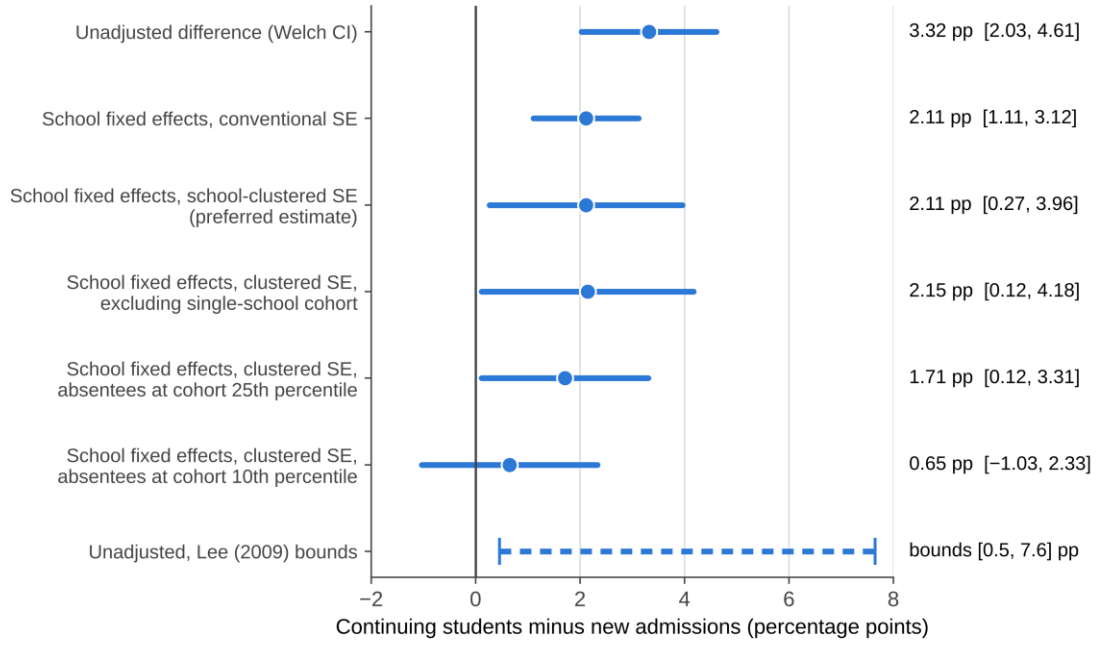



**Figure 3.** The estimated cohort difference across specifications. Markers show point estimates with 95% confidence intervals: Welch for the unadjusted difference, conventional or school-clustered for school fixed-

effects estimates, as labelled. The dashed bar shows Lee (2009) bounds on the unadjusted difference. Values on the right give the estimate and interval in percentage points (pp).

### Checks on Test Familiarity (RQ4)

Two checks bear on test familiarity, although neither can separate it from programme exposure. First, no item showed moderate or large differential item functioning. All 22 items fell in the ETS negligible category (A), with absolute Mantel–Haenszel D-DIF values below 1.0 (Appendix A, Table A1). The largest value was for Item 2 (identity, physics; D-DIF = −0.87), which favoured new admissions. The continuing students' advantage was thus spread across items rather than concentrated on a few. A general difference in reasoning would produce this pattern, but so would general familiarity with the test format. The DIF result therefore rules out only item-specific familiarity, such as recall of particular items.

Second, new admissions left more items unanswered or marked more than one option. Such responses made up 2.6% of their responses, against 1.1% for continuing students, and 19.3% of new admissions gave at least one, against 9.2% of continuing students. This is consistent with lower familiarity with the format. Its direct effect on the estimate was small. Crediting every unanswered or invalid response with the chance value of 0.25 changed the school-adjusted difference from 2.11 to 2.03 pp, 95% CI [0.19, 3.87]. Familiarity effects that work through faster or more confident reasoning on answered items cannot be detected in these data.

### Instrument Quality and the Competency Results (RQ4)

The weak items bear directly on the competency findings. The null result for identity rests mainly on one physics item (Item 2), on which new admissions scored 6.1 pp higher and whose discrimination was marginal (.17). Removing the weak identity item (Item 10) does not change the conclusion, as the raw gap on the remaining two items is −1.3 pp. The identity result may therefore reflect the particular items as much as any difference in reasoning. A second weak item measured conservation. A third measured classification and decentration and had a discrimination of zero, which suggests a keying error or a misleading distractor. Without that item, classification and decentration still showed the largest adjusted difference (+4.1 pp, 95% CI [0.1, 8.0]), so that difference is unlikely to be an artefact of a single item, although it then rests on two items. The strongest items, with point-biserial correlations of .40 to .45, measured transitivity, seriation and conservation.

## DISCUSSION

### A Small and Uncertain Observed Difference

Prior exposure to Kriya in Grades 6 and 7 was associated with a within-school difference of about 2 pp, or 0.11 SD, in reasoning at entry to Grade 8. With standard errors clustered by school, the 95% CI ran from 0.3 to 4.0 pp (0.01 to 0.21 SD). The estimate is below Cohen's (1988) threshold for a small effect. By benchmarks built from field trials with achievement outcomes it would fall in the medium range (Kraft, 2020), although those benchmarks come from causal designs and this one is not. It is far smaller than the effects reported in early CASE studies (Adey & Shayer, 1990; Shayer & Adey, 1993) and larger than the null result when a CASE successor was scaled up in England (Hanley et al., 2016). It is of a similar order to the effects that remained a year after remedial programmes in India ended (Banerjee et al., 2007), although outcomes and designs differ. None of these comparisons implies that the difference observed here was caused by Kriya.

Two features of the setting matter for interpretation. First, the outcome is reasoning operations, not curricular content. These operations develop slowly and are hard to teach to a test (Shayer & Ginsburg, 2009). Second, the measure was taken at the start of a new school year, after the annual vacation. Many education interventions fade once support is removed (Bailey et al., 2017, 2020). If the difference reflects the programme, it reflects what remained after the vacation. It may equally reflect familiarity with the test or differences between feeder schools, as discussed below. The estimate should be read with its uncertainty in view. The unadjusted gap of 3.3 pp overstated the within-school difference because continuing students

were concentrated in stronger schools. Clustering students within schools nearly doubled the standard error. Differential absence may inflate the estimate further: under the most pessimistic assumption, the school-adjusted difference fell to 0.7 pp and its interval included zero. The within-school estimate of about 2 pp is our best estimate, but the true difference could be close to zero. This is the study's main methodological finding. Each step from a raw comparison towards a more defensible one made the apparent advantage smaller or less certain.

**Competency-Level Differences: Observation and Interpretation**

We separate what was observed from how it might be interpreted. The observed finding is that continuing students' school-adjusted scores were higher on six of the seven competency measures. Point estimates were largest for classification and decentration (+3.8 pp) and transitivity (+3.4 pp) and smallest for generalisation (+0.9 pp) and identity (−0.9 pp). None of these differences survived adjustment for multiple comparisons, the subscales had reliabilities between .03 and .45, and each rested on three or four items. One interpretation is that, if the larger differences are real, they reflect greater practice of these operations in Kriya activities. Sorting objects by several attributes and inferring a relation between two objects through a third can be enacted directly with materials, and experiential activities offer repeated practice in these acts. This interpretation remains a hypothesis for future testing, not a finding of this study.

No difference was detected for generalisation, which was among the two hardest competencies for both groups. Generalisation requires students to extract a rule from particular cases and apply it to new ones, which is closer to formal than to concrete reasoning. The CASE literature suggests that this step depends less on the concrete activity than on what follows it: cognitive conflict, metacognitive talk and explicit bridging to new contexts (Adey & Shayer, 1994; Venville & Oliver, 2015). The finding that guided discovery outperforms unguided discovery points the same way (Alfieri et al., 2011; Lazonder & Harmsen, 2016). Kriya activities do include an investigation phase that asks students to extend what they observed, for example by testing other fruits in the voltaic cell activity (Appendix B). If the null result reflects the programme rather than the measure, it may reflect how consistently this phase is carried out in class. The generalisation items all discriminated adequately (Appendix A), but the subscale's reliability (.34) was still far too low to rule out a modest difference. Strengthening and monitoring the investigation and conclusion phases, where the experiential cycle moves from experience to abstraction (Kolb, 1984), is a design change worth testing. The null result for identity calls for still more caution. The subscale's reliability was close to zero (.03), and the result should not be read as evidence that the programme fails on identity until the items are revised.

The subject pattern is also descriptive. Chemistry was the weakest strand for both groups and showed the smallest difference, which within schools was close to zero (0.8 pp). All three weak items were chemistry items. Without them, chemistry scores rise to about the level of the other strands (41.9% and 39.8%), and the raw gap stays at about 2 pp. The low level of chemistry scores is therefore largely a measurement issue. Many chemistry ideas concern particles and changes that cannot be seen or handled directly, and they are harder to make concrete than ordering, sorting or comparing. Whether this explains the smaller difference cannot be determined from these data.

**Where in the Distribution the Difference Appears**

In raw terms, the difference lay in the lower and middle parts of the distribution: fewer continuing students scored at chance, and more reached the 50–75% band. Within schools, only the middle shift was clear. The floor difference largely reflected school composition. The floor is also the part of the distribution most exposed to differential absence and to familiarity with the format, since students unfamiliar with a test are more likely to guess or leave items blank. There was no difference at the top in either analysis. Raw zone-level gaps appeared larger where performance was lower, but within schools the zone differences were all positive, similar in size and not significantly different. We therefore describe a distributional pattern and do not infer that the programme benefits weaker students in particular. Whether concrete activity gives a first foothold in reasoning to students who lack one, while higher performance needs more abstract reasoning, is a question for a design with a pre-test. Absolute performance also matters. Both cohorts averaged close to

40%, and one in five continuing students still scored at chance. National survey results, although based on a different instrument, point the same way: average Grade 9 science performance is about 40% nationally and 37% in Karnataka (NCERT, 2025a, 2025b). Whatever the source of the difference between the groups, both remain far from mastery.

**Why Differences Vary Across Schools**

School membership explained about seventy times as much variance as cohort membership. Continuing students outscored new admissions in 63% of schools, a rate not significantly above an even split, and two schools in the same network produced opposite results. This variation is also why school-clustered standard errors were so much larger than conventional ones. Such variation is common in multi-site programmes and is often linked to implementation (Durlak & DuPre, 2008). Without fidelity data, we cannot test that explanation here (O'Donnell, 2008). The site-level contrasts identified in this study offer a starting point. The larger raw gaps in private and CBSE schools also need careful reading. Once school was held constant, they either shrank or became too uncertain to interpret. Where differences remain, they may reflect implementation, more stable enrolment, or who the new admissions were: students who move into private schools at Grade 8 may differ from those who move into government schools.

**Principal Alternative Explanations: Test Familiarity and Feeder-School Selection**

Test familiarity and feeder-school selection are the two principal alternative explanations for the observed difference, and either could account for much of it. Continuing students had sat the programme's assessments several times in Grades 6 and 7, whereas new admissions met the format for the first time. A meta-analysis found retest gains of about a third of a standard deviation between a first and second administration, rising to about half a standard deviation by the third, with larger gains when the same form is reused (Scharfen et al., 2018). The observed 0.11 SD difference could therefore partly, or even wholly, reflect prior exposure to the assessment itself. Our checks constrain only some forms of familiarity. The absence of DIF argues against item-specific recall. The higher rate of unanswered responses among new admissions shows a familiarity difference, but its direct effect on the estimate was small. General familiarity with the format, the item style and the testing routine would raise scores on all items alike and cannot be detected by DIF.

The second explanation is feeder-school selection. School fixed effects control for differences between receiving secondary schools, but they do not eliminate individual-level or feeder-school-level selection differences between continuing and newly admitted students. Continuing students came from higher primary schools that ran Kriya, and new admissions from schools that did not. If the schools that joined Kriya were stronger, better resourced or served more advantaged families, the estimate overstates any programme contribution. If they were selected because they were weaker, it understates it. The records contain neither feeder-school identifiers nor student background data, so this threat cannot be tested here. Recording each student's feeder school and basic background variables would allow future analyses to include feeder-school fixed effects or matching.

**Implications for Programme Evaluation**

The study suggests seven practices for evaluating school programmes that rely on naturally occurring comparison groups.

First, report within-school estimates alongside raw gaps. Here, school composition accounted for a third of the raw difference. Programme participants are rarely spread evenly across schools, so raw comparisons will often mislead.

Second, cluster standard errors at the level at which students are grouped. Here, clustering by school nearly doubled the standard error, and an interval that appeared comfortably above zero moved close to it. With 60 schools, student-level standard errors would have overstated the precision of the estimate.

Third, record attendance and enrolment status, with reasons, for both groups. The near threefold difference in absence was a major threat to inference in this study. Part of it may not be absence at all. Continuing students can remain on the roll after they have left in practice, whereas new admissions are, by definition, recently enrolled. Recording why students are missing would help separate absence from undocumented exit.

Fourth, publish sensitivity analyses. Showing how an estimate changes under plausible assumptions about missing students is less striking than a single figure of 3.3 pp, but it is more defensible and more useful for decisions. Formal approaches, such as trimming bounds (Lee, 2009) and tipping-point analyses, strengthen this further.

Fifth, check item functioning and subscale reliability before interpreting competency-level results, and adjust for multiple comparisons. A competency measured by three items, one of which does not discriminate, cannot separate a programme failure from a measurement failure.

Sixth, treat programme monitoring tests as evaluation instruments. When such tests are used to estimate differences between groups, they need the scrutiny given to any outcome measure: scoring from raw responses, item analysis and reliability reporting. In this study, re-scoring from raw responses alone corrected 84 records.

Seventh, account for test familiarity when programme students have been assessed before and comparison students have not, for example by using alternate forms linked by common anchor items, by giving the comparison group a practice form, or by comparing the groups again at a later administration.

### Implications for Policy and Practice

National policy now asks for experiential learning as standard pedagogy (MHRD, 2020). The present findings are consistent with sustained hands-on activity being associated with a modest difference in reasoning, but they cannot show that it causes one, and the difference varies across schools. Three implications follow for programme designers and state systems. First, if the aim is reasoning that generalises, experiential sessions should be designed to end in reflection and transfer, not only in activity, as the wider literature suggests (Adey & Shayer, 1994; Alfieri et al., 2011). Second, support should be targeted at schools where programme students show no advantage, since school differences dwarf cohort differences. Third, competency-based assessment at state and national level could include short reasoning scales, built with enough items per competency to be reliable. This would allow systems to monitor the outcomes that experiential pedagogy is meant to improve, rather than content knowledge alone. The present data also create an opportunity. Because Kriya secondary schools run the programme from Grade 8, new admissions begin the programme this year. If both groups are assessed again at the end of Grade 8, this baseline becomes a common pre-test, giving a pretest-posttest comparison of the two groups. Growth comparisons would show whether the gap narrows once both groups have received the programme and know the test format. A familiarity advantage should shrink quickly, while a difference in reasoning should not.

## LIMITATIONS

The design is a non-equivalent comparison-group design without a pre-test and does not support causal claims. Its two principal threats, feeder-school selection and test familiarity, are discussed above. Students were not randomly assigned, and new admissions may differ from continuing students in prior schooling, home background or motivation. The direction of this bias is unknown. New admissions who move from private schools may be advantaged, while those who move after migration may be disadvantaged. We had no student-level background data with which to adjust, and no record of each student's feeder school. There was no common pre-test, so we estimate a difference in level at Grade 8, not in growth over Grades 6 and 7. The attendance analysis covered absent students but not the 330 who had transferred out. The direction of any resulting bias is unknown. If stronger students left for private schools, the observed gap would understate the programme difference; if weaker or migrant students left, it would overstate it. In the 25th-percentile scenario, absent students were imputed at their own cohort's percentile, which carries part of the

observed advantage into the imputed scores. Standard errors are clustered by school, but with 60 schools, and fewer within subgroups, cluster-robust inference is itself approximate; the wild cluster bootstrap gave a similar result for the main estimate. The competency results rest on subscales of three or four items with reliabilities between .03 and .45. No confirmatory factor analysis or item response model was fitted, so the seven competencies cannot be treated as distinct, reliably measured constructs. Several subgroup analyses involve few schools. Paper-based multiple-choice items may also under-represent what students learn through hands-on work (Chi et al., 2021), and the test measures reasoning operations in science contexts rather than the full range of scientific reasoning skills, such as designing controlled experiments (Zimmerman, 2007). Finally, the cohort analysis relied on the year encoded in identifiers, and the two smaller cohorts were each concentrated in one or two schools.

### Future Research

Five lines of work follow. The most valuable is a pretest-posttest longitudinal design in which continuing students and new admissions are assessed at the beginning and end of Grade 8, with this baseline as the pre-test. Second, the familiarity explanation can be tested directly by using alternate test forms linked by common anchor items, by including an assessment unfamiliar to both groups, and by repeating the DIF analysis at the end-of-year administration, when both groups have met the format. Third, recording each student's feeder school and basic background variables would allow feeder-school selection to be modelled. Fourth, the instrument needs more items per competency, with competency-level reliability estimates, item response modelling and, where appropriate, confirmatory factor analysis before competency profiles are interpreted. Fifth, fidelity measures linked to school outcomes would test whether implementation explains the variation between schools, and a design experiment that adds explicit bridging and metacognitive discussion to Kriya activities would test whether generalisation can be moved.

## CONCLUSION

Students who took part in Kriya during Grades 6 and 7 entered Grade 8 with a small observed advantage in reasoning over newly admitted classmates. Within schools, the difference was about 2 pp, or 0.11 SD, with a school-clustered 95% CI of 0.3 to 4.0 pp. It was sensitive to differential absence, may partly reflect familiarity with the test or differences between feeder schools, and could not be reliably located in particular competencies. Both groups remained far from mastery. The study's clearest contribution is methodological. A raw difference of 3.3 pp became 2.1 pp once school composition was considered, became much less precise once the clustering of students in schools was recognised, and fell below 1 pp under pessimistic assumptions about absent students. For evaluators who rely on naturally occurring comparison groups, reporting this sequence is more useful and more defensible than reporting a single number.

### Conflict of Interest

The authors are employed by Prayoga Institute of Education Research, which designs and implements the Kriya programme. The analysis reported here was carried out by the organisation's research team. The authors declare no other conflicts of interest.

### Data Availability Statement

Anonymised data and analysis code are available from the corresponding author on reasonable request, subject to approval by Prayoga Institute of Education Research and the relevant education authorities.

### Funding

This study was not funded by any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

## ACKNOWLEDGEMENTS

The authors thank the headteachers, facilitators, academic coordinators of Prayoga Institute of Education Research and students of the participating schools, the data management team who prepared the assessment

records and the management of Prayoga. We extend a special thanks to Dr. K S Nagabhushana, Director-Research, Prayoga for guiding us.

## REFERENCES

1. Adey, P. (2005). Issues arising from the long-term evaluation of cognitive acceleration programs. Research in Science Education, 35(1), 3–22. https://doi.org/10.1007/s11165-004-3430-5
2. Adey, P., & Shayer, M. (1990). Accelerating the development of formal thinking in middle and high school students. Journal of Research in Science Teaching, 27(3), 267–285. https://doi.org/10.1002/tea.3660270309
3. Adey, P., & Shayer, M. (1994). Really raising standards: Cognitive intervention and academic achievement. Routledge.
4. Alfieri, L., Brooks, P. J., Aldrich, N. J., & Tenenbaum, H. R. (2011). Does discovery-based instruction enhance learning? Journal of Educational Psychology, 103(1), 1–18. https://doi.org/10.1037/a0021017
5. Bailey, D., Duncan, G. J., Odgers, C. L., & Yu, W. (2017). Persistence and fadeout in the impacts of child and adolescent interventions. Journal of Research on Educational Effectiveness, 10(1), 7–39. https://doi.org/10.1080/19345747.2016.1232459
6. Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and fade-out of educational-intervention effects: Mechanisms and potential solutions. Psychological Science in the Public Interest, 21(2), 55–97. https://doi.org/10.1177/1529100620915848
7. Banerjee, A. V., Cole, S., Duflo, E., & Linden, L. (2007). Remedying education: Evidence from two randomized experiments in India. The Quarterly Journal of Economics, 122(3), 1235–1264. https://doi.org/10.1162/qjec.122.3.1235
8. Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1), 289–300. https://doi.org/10.1111/j.2517-6161.1995.tb02031.x
9. Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-based improvements for inference with clustered errors. The Review of Economics and Statistics, 90(3), 414–427. https://doi.org/10.1162/rest.90.3.414
10. Cameron, A. C., & Miller, D. L. (2015). A practitioner's guide to cluster-robust inference. Journal of Human Resources, 50(2), 317–372. https://doi.org/10.3368/jhr.50.2.317
11. Cheung, A. C. K., & Slavin, R. E. (2016). How methodological features affect effect sizes in education. Educational Researcher, 45(5), 283–292. https://doi.org/10.3102/0013189X16656615
12. Chi, S., Liu, X., & Wang, Z. (2021). Comparing student science performance between hands-on and traditional item types: A many-facet Rasch analysis. Studies in Educational Evaluation, 70, Article 100998. https://doi.org/10.1016/j.stueduc.2021.100998
13. Cohen, J. (1988). Statistical power analysis for the behavioral sciences (2nd ed.). Lawrence Erlbaum Associates.
14. Conn, K. M. (2017). Identifying effective education interventions in sub-Saharan Africa: A meta-analysis of impact evaluations. Review of Educational Research, 87(5), 863–898. https://doi.org/10.3102/0034654317712025
15. Corning, A., Broda, M. D., Lucas, B.-L., Becker, J. D., & Bae, C. L. (2023). An inclusive school for computer science: Evaluating early impact with propensity score matching. Studies in Educational Evaluation, 79, Article 101293. https://doi.org/10.1016/j.stueduc.2023.101293
16. Durlak, J. A., & DuPre, E. P. (2008). Implementation matters: A review of research on the influence of implementation on program outcomes and the factors affecting implementation. American Journal of Community Psychology, 41(3–4), 327–350. https://doi.org/10.1007/s10464-008-9165-0
17. Ebel, R. L., & Frisbie, D. A. (1991). Essentials of educational measurement (5th ed.). Prentice Hall.
18. Feldt, L. S. (1965). The approximate sampling distribution of Kuder–Richardson reliability coefficient twenty. Psychometrika, 30(3), 357–370. https://doi.org/10.1007/BF02289499
19. Furtak, E. M., Seidel, T., Iverson, H., & Briggs, D. C. (2012). Experimental and quasi-experimental studies of inquiry-based science teaching: A meta-analysis. Review of Educational Research, 82(3), 300–329. https://doi.org/10.3102/0034654312457206

20. Gottfried, M. A. (2010). Evaluating the relationship between student attendance and achievement in urban elementary and middle schools: An instrumental variables approach. American Educational Research Journal, 47(2), 434–465. https://doi.org/10.3102/0002831209350494
21. Hanley, P., Böhnke, J. R., Slavin, R., Elliott, L., & Croudace, T. (2016). Let's Think Secondary Science: Evaluation report and executive summary. Education Endowment Foundation. https://educationendowmentfoundation.org.uk/projects-and-evaluation/projects/lets-think-secondary-science
22. Hedges, L. V., & Hedberg, E. C. (2007). Intraclass correlation values for planning group-randomized trials in education. Educational Evaluation and Policy Analysis, 29(1), 60–87. https://doi.org/10.3102/0162373707299706
23. Hill, C. J., Bloom, H. S., Black, A. R., & Lipsey, M. W. (2008). Empirical benchmarks for interpreting effect sizes in research. Child Development Perspectives, 2(3), 172–177. https://doi.org/10.1111/j.1750-8606.2008.00061.x
24. Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel–Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), Test validity (pp. 129–145). Lawrence Erlbaum Associates.
25. Holm, S. (1979). A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics, 6(2), 65–70.
26. Inhelder, B., & Piaget, J. (1958). The growth of logical thinking from childhood to adolescence: An essay on the construction of formal operational structures (A. Parsons & S. Milgram, Trans.). Basic Books.
27. Inhelder, B., & Piaget, J. (1964). The early growth of logic in the child: Classification and seriation (E. A. Lunzer & D. Papert, Trans.). Routledge & Kegan Paul.
28. Kolb, D. A. (1984). Experiential learning: Experience as the source of learning and development. Prentice-Hall.
29. Kraft, M. A. (2020). Interpreting effect sizes of education interventions. Educational Researcher, 49(4), 241–253. https://doi.org/10.3102/0013189X20912798
30. Kuder, G. F., & Richardson, M. W. (1937). The theory of the estimation of test reliability. Psychometrika, 2(3), 151–160. https://doi.org/10.1007/BF02288391
31. Lawson, A. E. (1978). The development and validation of a classroom test of formal reasoning. Journal of Research in Science Teaching, 15(1), 11–24. https://doi.org/10.1002/tea.3660150103
32. Lazonder, A. W., & Harmsen, R. (2016). Meta-analysis of inquiry-based learning: Effects of guidance. Review of Educational Research, 86(3), 681–718. https://doi.org/10.3102/0034654315627366
33. Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. The Review of Economic Studies, 76(3), 1071–1102. https://doi.org/10.1111/j.1467-937X.2009.00536.x
34. McGraw, K. O., & Wong, S. P. (1992). A common language effect size statistic. Psychological Bulletin, 111(2), 361–365. https://doi.org/10.1037/0033-2909.111.2.361
35. Ministry of Human Resource Development. (2020). National education policy 2020. Government of India. https://www.education.gov.in/sites/upload_files/mhrd/files/NEP_Final_English.pdf
36. Minner, D. D., Levy, A. J., & Century, J. (2010). Inquiry-based science instruction—what is it and does it matter? Results from a research synthesis years 1984 to 2002. Journal of Research in Science Teaching, 47(4), 474–496. https://doi.org/10.1002/tea.20347
37. Muralidharan, K., Singh, A., & Ganimian, A. J. (2019). Disrupting education? Experimental evidence on technology-aided instruction in India. American Economic Review, 109(4), 1426–1460. https://doi.org/10.1257/aer.20171112
38. National Council of Educational Research and Training. (2023). National curriculum framework for school education 2023.
39. National Council of Educational Research and Training. (2025a). PARAKH Rashtriya Sarvekshan 2024: National report. https://parakh.ncert.gov.in/sites/default/files/2025-07/REPORT_India_IND.pdf
40. National Council of Educational Research and Training. (2025b). PARAKH Rashtriya Sarvekshan 2024: Karnataka state report. https://parakh.ncert.gov.in/sites/default/files/2025-07/REPORT_Karnataka_IND029.pdf

41. O'Donnell, C. L. (2008). Defining, conceptualizing, and measuring fidelity of implementation and its relationship to outcomes in K–12 curriculum intervention research. Review of Educational Research, 78(1), 33–84. https://doi.org/10.3102/0034654307313793
42. Oliver, M., Venville, G., & Adey, P. (2012). Effects of a cognitive acceleration programme in a low socioeconomic high school in regional Australia. International Journal of Science Education, 34(9), 1393–1410. https://doi.org/10.1080/09500693.2012.673241
43. Piaget, J., & Inhelder, B. (1969). The psychology of the child (H. Weaver, Trans.). Basic Books.
44. Scharfen, J., Peters, J. M., & Holling, H. (2018). Retest effects in cognitive ability tests: A meta-analysis. Intelligence, 67, 44–66. https://doi.org/10.1016/j.intell.2018.01.003
45. Shadish, W. R., Cook, T. D., & Campbell, D. T. (2002). Experimental and quasi-experimental designs for generalized causal inference. Houghton Mifflin.
46. Shayer, M. (1999). Cognitive acceleration through science education II: Its effects and scope. International Journal of Science Education, 21(8), 883–902. https://doi.org/10.1080/095006999290345
47. Shayer, M., & Adey, P. S. (1993). Accelerating the development of formal thinking in middle and high school students IV: Three years after a two-year intervention. Journal of Research in Science Teaching, 30(4), 351–366. https://doi.org/10.1002/tea.3660300404
48. Shayer, M., & Ginsburg, D. (2009). Thirty years on – a large anti-Flynn effect? (II): 13- and 14-year-olds. Piagetian tests of formal operations norms 1976–2006/7. British Journal of Educational Psychology, 79(3), 409–418. https://doi.org/10.1348/978185408X383123
49. Shayer, M., Küchemann, D. E., & Wylam, H. (1976). The distribution of Piagetian stages of thinking in British middle and secondary school children. British Journal of Educational Psychology, 46(2), 164–173. https://doi.org/10.1111/j.2044-8279.1976.tb02308.x
50. Venville, G., & Oliver, M. (2015). The impact of a cognitive acceleration programme in science on students in an academically selective high school. Thinking Skills and Creativity, 15, 48–60. https://doi.org/10.1016/j.tsc.2014.11.004
51. Welch, B. L. (1947). The generalization of 'Student's' problem when several different population variances are involved. Biometrika, 34(1–2), 28–35. https://doi.org/10.1093/biomet/34.1-2.28
52. Zimmerman, C. (2007). The development of scientific thinking skills in elementary and middle school. Developmental Review, 27(2), 172–223. https://doi.org/10.1016/j.dr.2006.12.001
53. Zwick, R. (2012). A review of ETS differential item functioning assessment procedures: Flagging rules, minimum sample size requirements, and criterion refinement (Research Report No. RR-12-08). Educational Testing Service. https://doi.org/10.1002/j.2333-8504.2012.tb02290.x

## APPENDIX A. ITEM STATISTICS

**Table A1. Item statistics by cohort**

| Item | Strand | Competency | Continuing (%) | New admissions (%) | Gap (pp) | rpb | MH D-DIF | ETS class |
|---|---|---|---|---|---|---|---|---|
| 1 | Physics | Classification and decentration | 40.2 | 32.6 | 7.6 | 0.26 | 0.44 | A |
| 2 | Physics | Identity | 44.7 | 50.8 | −6.1 | 0.17 | −0.87 | A |
| 3 | Physics | Seriation | 53.7 | 47.6 | 6.1 | 0.3 | 0.24 | A |
| 4 | Physics | Transitivity | 44.7 | 39.4 | 5.3 | 0.25 | 0.25 | A |
| 5 | Physics | Inductive logic | 56.3 | 50.9 | 5.4 | 0.37 | 0.1 | A |
| 6 | Physics | Conservation | 45.5 | 42.6 | 2.8 | 0.25 | −0.02 | A |
| 7 | Physics | Generalisation | 27.4 | 25.8 | 1.6 | 0.25 | −0.12 | A |
| 8 | Chemistry | Generalisation | 36.6 | 38.7 | −2.1 | 0.21 | −0.51 | A |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 9 | Chemistry | Classification and decentration | 24.1 | 21.8 | 2.3 | 0 | 0.21 | A |
| 10 | Chemistry | Identity | 35.7 | 33.3 | 2.4 | 0.07 | 0.06 | A |
| 11 | Chemistry | Seriation | 50.4 | 46.6 | 3.8 | 0.4 | −0.14 | A |
| 12 | Chemistry | Transitivity | 39.1 | 35.5 | 3.6 | 0.34 | −0.02 | A |
| 13 | Chemistry | Inductive logic | 45.2 | 43.4 | 1.8 | 0.32 | −0.25 | A |
| 14 | Chemistry | Conservation | 27.5 | 26.7 | 0.7 | 0.08 | 0.02 | A |
| 15 | Chemistry | Generalisation | 38.1 | 34.9 | 3.3 | 0.21 | 0.09 | A |
| 16 | Biology | Classification and decentration | 38.3 | 34.1 | 4.2 | 0.29 | 0.15 | A |
| 17 | Biology | Identity | 41.5 | 37.9 | 3.5 | 0.35 | −0.04 | A |
| 18 | Biology | Seriation | 50.3 | 45.2 | 5.1 | 0.37 | 0.04 | A |
| 19 | Biology | Transitivity | 47.9 | 41.5 | 6.4 | 0.45 | 0.15 | A |
| 20 | Biology | Inductive logic | 40.7 | 34.2 | 6.5 | 0.33 | 0.34 | A |
| 21 | Biology | Conservation | 36.9 | 31.9 | 5.1 | 0.4 | 0.1 | A |
| 22 | Biology | Generalisation | 35.6 | 32 | 3.7 | 0.37 | −0.03 | A |

Note. $r_{pb}$ = point-biserial correlation between the item and the total score excluding that item. Items 9, 10 and 14 fall far below the commonly used minimum of .20; Item 2 is marginal. MH D-DIF = Mantel–Haenszel delta difference, with new admissions as the reference group and students matched on total score; negative values favour new admissions. Class A = negligible DIF (|D-DIF| < 1.0 or not significant); no item reached class B or C, before or after removing flagged items from the matching score. Gaps were computed from unrounded percentages. pp = percentage points.

## APPENDIX B. SAMPLE KRIYA ACTIVITIES

Tables B1 and B2 summarise two activities from the Kriya course books. Both follow the programme's four-phase structure of orientation, conceptualisation, investigation and conclusion.

**Table B1. Sample activity 1: What makes moving objects stop? (friction)**

| **Element** | **Description** |
|---|---|
| **Learning objective** | Define friction. |
| **Competency** | Inquiry |
| **Suggested time** | 20 minutes, group activity, inquiry-based learning. |
| **Materials** | Inclined plane set-up, cylindrical object, laminate sheet, cardboard and sandpaper. |
| **Orientation** | A rolling ball, a pushed toy car and a bus with its engine switched off all slow down and stop. Students are asked why moving objects stop by themselves. |

| | |
|---|---|
| **Conceptualisation** | Students release a cylinder down the inclined plane onto a laminate sheet, then cardboard, then sandpaper, and record the distance it travels on each surface (Figure B1). |
| **Investigation** | Students state what the three cases share (contact between surfaces; the cylinder stops), infer an opposing force, explain why the distances differ, and list the surfaces in order of increasing friction. |
| **Conclusion** | Friction is a contact force that opposes motion. Its size depends on the nature of the surfaces in contact. |

Note. Summarised from the Kriya course book. The final investigation step, ordering the surfaces by friction, is a seriation task.

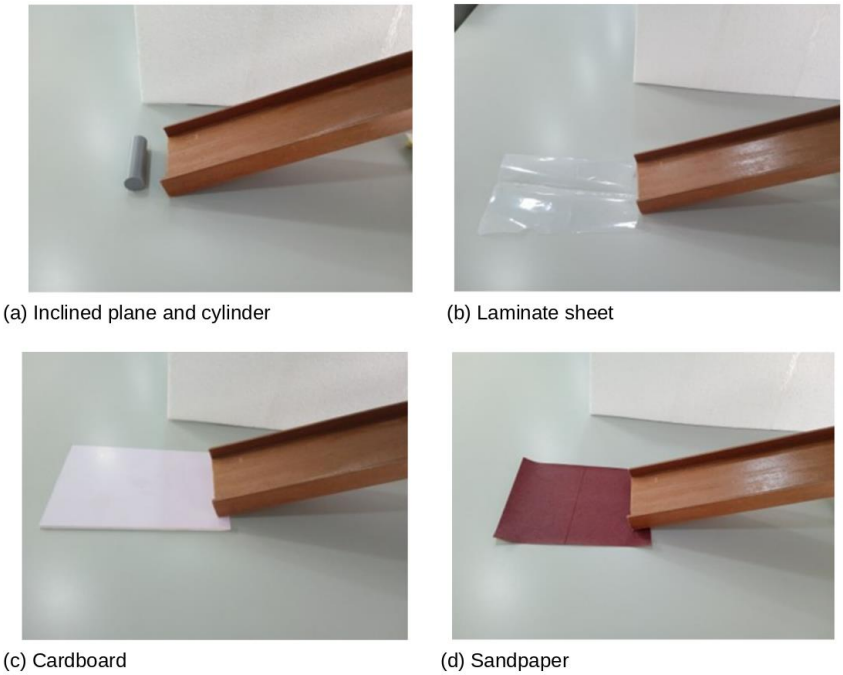


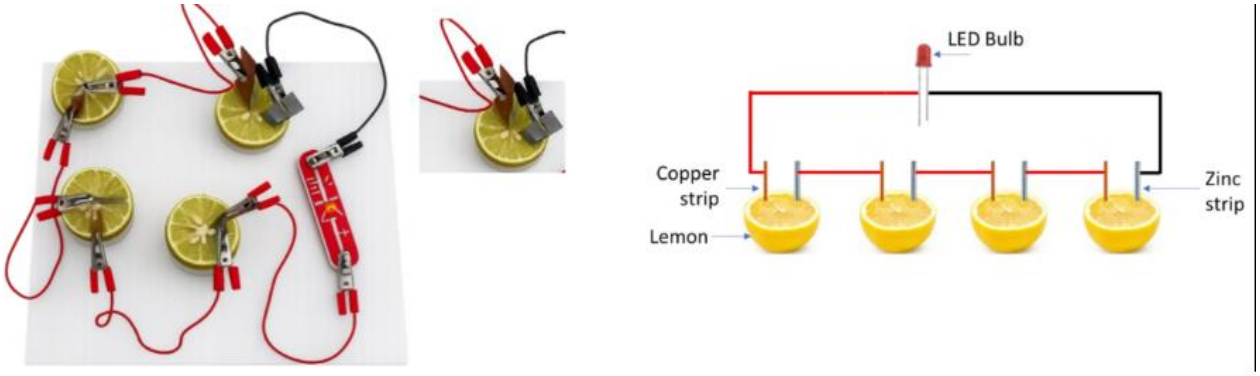

**Figure B1.** Set-up for the friction activity. Students release the cylinder down the inclined plane onto each surface in turn and record the distance it travels.

**Table B2.** Sample activity 2: Can a lemon light a bulb? (voltaic cell)

| Element | Description |
|---|---|
| **Learning objectives** | Construct a simple voltaic cell; describe its parts; explain how it works. |
| **Competency** | Identity, Seriation, Inquiry |
| **Materials** | Two lemons, a knife, connecting wires, zinc and copper strips (3.5 cm × 1 cm), an LED and double-sided tape. |
| **Orientation** | Two children find a note saying that a bulb can be lit without a battery, because "electricity is hidden in everyday things". |
| **Conceptualisation** | Students cut two lemons in half, insert a zinc and a copper strip into each half, connect the four halves in series to an LED and observe it glow (Figure B2). |

| | |
|---|---|
| **Investigation** | Students explain how the LED glows without a battery and identify the electrodes and the electrolyte. They then repeat the activity with other low-cost fruits or vegetables, such as tomato, potato, tamarind pulp, orange or onion, and list those that work. |
| **Conclusion** | A voltaic cell produces an electric current through a chemical reaction between two different electrodes and an electrolyte. The electrochemical series shows which metal loses electrons more readily. |

Note. Summarised from the Kriya course book. Repeating the activity with other materials asks students to test whether the finding generalises.




(a) Lemon cells lighting an LED (b) Circuit diagram

**Figure B2.** The lemon voltaic cell. Four lemon halves, each holding a copper and a zinc strip, are connected in series to light an LED.