International Journal of Science Education and Teaching
Vol. 5, No. 3, pp. 128-144, Sep. – Dec. 2026
http://doi.org/10.14456/ijset.2026.09

Research Article

When Does Inquiry Help? Nonlinear Science-Instruction Associations and Changing Achievement Contexts in Southeast Asian PISA Data, 2015-2022

MD Showket Hasan[1*]
[1]*Department of Science, Science Math Bilingual Program, Chalermkwansatree School, Phitsanulok 65000, Thailand*
*Email: showket.h@chs.ac.th

Received: 9 Aug. 2026 Revised: 1 Nov. 2026 Accepted: 11 Nov. 2026

Abstract. This study examined when more frequent inquiry-oriented science instruction is associated with higher science achievement, and whether that relationship travels across education systems. We analysed PISA 2015 student data from Thailand, whose complete-case analytic sample comprised n = 7,250 students (estimated weighted population 545,378). Singapore, Indonesia and Viet Nam served as prespecified replication systems, and Malaysia as a sensitivity case. The instructional results were then situated within 2015-2022 changes in achievement and socioeconomic gradients in Thailand, Singapore and Indonesia. Survey-weighted models pooled all 10 science plausible values and accounted for complex sampling using 80 BRR-Fay replicate weights. Restricted cubic splines showed a robust nonlinear association in Thailand (Wald $\chi^2(2) = 37.57$, $p < .001$); the finding survived trimmed-support and response-propensity sensitivities. The shape varied with teacher-directed instruction, teacher support and socioeconomic status, but not science self-efficacy after false-discovery-rate correction. Nonlinearity replicated in Singapore and Indonesia, but not clearly in Viet Nam; common-support curves did not indicate a universal threshold. Thailand's mean science score was 16.55 points lower in 2022 than in 2018, whereas its socioeconomic gradient showed no clear 2015-2022 change. Results support a conditional, guidance-sensitive account of inquiry; they also underscore that cross-sectional PISA associations do not identify instructional effects or explain later national trends. For practice, these results are consistent with prior guidance research in suggesting that the configuration of inquiry, rather than its reported frequency, is the more useful design focus; sequencing, scaffolding and consolidation were not measured in these data, so this study cannot estimate what they would add.

Keywords: inquiry-based science education; PISA; science achievement; instructional guidance; educational inequality

INTRODUCTION

Inquiry is central to contemporary science education because it asks learners to formulate questions, work with evidence, explain phenomena and engage in practices resembling scientific knowledge building. Throughout this article, inquiry-oriented science instruction is operationalised through the PISA 2015 IBTEACH index: a student-reported summary of how frequently nine inquiry-related practices occur in science lessons: students explain their ideas; students spend time in the laboratory doing practical experiments; students are required to argue about science questions; students draw conclusions from an experiment they conducted; the teacher explains how a science

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

idea can be applied; students design their own experiments; a class debate about investigations takes place; the teacher explains the relevance of broad science concepts to students' lives; and students do an investigation to test ideas (OECD, 2016b). Two of the nine items (ST098Q06TA and ST098Q09TA) describe teacher explanation rather than student activity, so the index is not a pure measure of student-led investigation (Table S9). Yet the empirical record does not support a simple claim that increasing the frequency of inquiry activities will invariably raise achievement. Experimental and quasi-experimental syntheses generally report positive average effects, particularly when inquiry is supported by explicit guidance (Furtak et al., 2012; Lazonder & Harmsen, 2016; Minner et al., 2010). At the same time, secondary analyses of international large-scale assessments frequently recover negative, mixed or curvilinear associations between student-reported inquiry frequency and science performance (Cairns, 2019; Jerrim et al., 2019; Jiang & McComas, 2015; Oliver et al., 2021; Teig et al., 2018). Two studies fix the contrast. Pooling PISA 2015 data across 69 education systems, Cairns (2019) found the composite inquiry index negatively associated with science achievement (approximately −6.5 score points per index unit); the item-level analyses beneath that average were sharply heterogeneous, ranging from roughly −11 points for designing one's own experiments to about +12 points for teacher explanations of how science ideas apply, with two practices showing inverted-U patterns that peaked when they occurred in "some" or "most" lessons rather than in every lesson. Analysing Norwegian TIMSS 2015 data at the classroom level, Teig et al. (2018) likewise documented an inverted-U relationship: achievement rose alongside inquiry-based teaching up to moderate frequencies and fell at the highest frequencies. The apparent contradiction matters for theory and practice. It may arise not because one evidence base is wrong, but because the label 'inquiry' spans instructional arrangements that differ in guidance, cognitive demand, frequency, implementation quality and learner readiness.

Theoretical arguments on guidance make this heterogeneity plausible. Minimal guidance can overload novices when domain knowledge and self-regulatory resources are insufficient (Kirschner et al., 2006), whereas scaffolds can preserve learners' epistemic agency while reducing unproductive search (Hmelo-Silver et al., 2007). Inquiry therefore need not be opposed to teacher direction. It can be organised through prompts, modelling, feedback and sequencing that make evidence-based reasoning tractable. Empirical work using PISA 2015 has similarly distinguished guided from independent forms of inquiry and shown that their relations with outcomes differ (Aditomo & Klieme, 2020). Analyses of high-performing systems and individual inquiry activities also indicate that aggregated indices may obscure the instructional configurations that students actually encounter (Cairns, 2019; Lau & Lam, 2017). Guidance is consequently best treated as a feature that may condition the inquiry-achievement relationship, not as the conceptual opposite of inquiry (Vorholzer & von Aufschnaiter, 2019).

A second problem concerns functional form. A single linear coefficient assumes that each one-unit increase in reported inquiry frequency carries the same expected achievement difference across the observed range. That assumption is stringent. Moderate exposure may create opportunities to reason with evidence, whereas very frequent activity without commensurate structure, curricular coherence or consolidation may produce diminishing or reversing associations. Teig et al. (2018) documented a curvilinear association in Norwegian TIMSS 2015 data, while Cairns (2019) showed that the frequency pattern differs across specific inquiry practices and that optima can sit at intermediate frequencies. A flexible nonlinear model is therefore more aligned with the theory than choosing in advance between a uniformly positive, uniformly negative or quadratic 'sweet spot' story. Crucially, however, a fitted turning region in observational data is not a prescriptive dosage. It can reflect selection, reporting, omitted variables or index construction as well as instruction.

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

A third problem is transportability. PISA-based findings vary across countries, and the measurement of broad teaching indices may not operate identically across cultural and educational contexts (Aditomo & Klieme, 2020; McConney et al., 2014; Oliver et al., 2021). Students interpret frequency categories within locally familiar classroom routines; systems differ in curriculum, teacher preparation, school composition and the balance between teacher explanation and student activity. Consequently, a shape estimated in one system cannot be assumed to define a regional threshold. Cross-system analysis should ask whether nonlinearity replicates over shared empirical support, while retaining system-specific levels and uncertainty. This is especially important in Southeast Asia, where neighbouring systems occupy markedly different achievement distributions and were treating the region as homogeneous would be analytically misleading. The present design therefore treats transportability as an empirical question rather than a default. Replication systems were prespecified to differ in achievement level, curriculum and classroom routine, so that a recurring need for nonlinear functional form would constitute regional evidence, while divergence would itself be informative about the limits of importing national curves. Because visual contrasts outside overlapping IBTEACH ranges would be driven by extrapolation, cross-system comparison is restricted to common empirical support. The appropriate policy question is thus not 'How much inquiry does Southeast Asia need?' but 'Which forms of inquiry, with what guidance, for which learners, under which system conditions?'

Finally, instructional associations are interpreted within changing educational environments. PISA 2015 provided the focal science and instructional measures, whereas later cycles permit repeated cross-sectional descriptions of science performance and socioeconomic gradients. These temporal results can show whether Thailand's broader performance and equity context changed relative to selected comparison systems. They cannot show that 2015 inquiry experiences caused later national outcomes, nor can three independent cross-sections identify a pandemic effect. Separating the within-cycle instructional layer from the cross-cycle contextual layer prevents an attractive but invalid causal narrative.

The present study contributes in four ways. First, it estimates Thailand's 2015 inquiry-achievement function using a prespecified sequence from linear to restricted cubic spline models. Second, it tests whether that function varies with science self-efficacy, perceived teacher support, teacher-directed instruction and socioeconomic status. Third, it examines replication in Singapore, Indonesia and Viet Nam over common inquiry support, with Malaysia retained as a sensitivity case—a design that operationalises the transportability concern argued above. Fourth, it situates those findings alongside 2015-2022 changes in science means and socioeconomic gradients in Thailand, Singapore and Indonesia. This design responds to calls for clearer conceptualisation, stronger large-scale-assessment methods and more circumspect reporting (Hopfenbeck et al., 2018; Rutkowski et al., 2010; Teig et al., 2022).

RESEARCH OBJECTIVES

RQ1. What is the functional form of the adjusted association between inquiry-based science instruction and science achievement in Thailand in PISA 2015?

RQ2. Does that association vary with science self-efficacy, perceived teacher support, teacher-directed instruction or socioeconomic status?

RQ3. Does the nonlinear association replicate in Singapore, Indonesia and Viet Nam over common empirical support, and is it robust when Malaysia is added as a sensitivity case?

RQ3's multi-system design exists precisely because transportability cannot be assumed: if nonlinearity recurs across systems with different achievement distributions, curricula and classroom routines, the need for flexible functional form is a regional property rather than a Thai idiosyncrasy, and if it does not, importing a single national curve would be unjustified.

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

RQ4. How did mean science achievement and socioeconomic gradients differ across the 2015, 2018 and 2022 PISA cross-sections in Thailand, Singapore and Indonesia?

RESEARCH METHODOLOGY

Design and data

We conducted a secondary analysis of anonymised public-use student data from the OECD Programme for International Student Assessment (PISA). PISA assesses nationally representative samples of 15-year-old students and reports science literacy through multiple plausible values rather than a single observed score (OECD, 2016a, 2017a, 2023; von Davier et al., 2009). The study followed a two-layer design.

Layer A: instructional associations (PISA 2015). Layer A used PISA 2015 because science was the major domain and the focal instructional indices were available. Thailand was the focal system; Singapore, Indonesia and Viet Nam were prespecified replication systems; Malaysia was a sensitivity-only case.

Layer B: repeated cross-sectional context (2015-2022). Layer B used the 2015, 2018 and 2022 independent cross-sections for Thailand, Singapore and Indonesia.

Country selection and roles. Thailand was the focal system because the motivating question concerns its instructional-achievement relationship and its changing performance context. Singapore, Indonesia and Viet Nam were prespecified as replication systems: they provide Southeast Asian contrasts in achievement level, curriculum structure and classroom routine while sharing regional proximity. Malaysia was prespecified as a 2015 sensitivity case only. Two boundary decisions require explanation. First, Viet Nam was excluded from the Layer B trend panel because OECD technical documentation attaches cross-cycle comparability and adjudication caveats to its successive cycles (OECD, 2023), and the locked protocol restricted trend systems to those whose cross-cycle comparability status met all criteria. Second, Malaysia was not elevated from sensitivity to a primary trend system: the full three-cycle panel required complete, adjudicated, trend-compatible measurements under the locked eligibility criteria, and Malaysia's available temporal data did not satisfy those panel requirements. Specifically, in PISA 2015 Malaysia's weighted school participation rate (51% initial, 52% after replacement) fell well below the 85% OECD technical reporting standard, leading OECD technical documentation to attach sampling coverage and response-rate adjudication caveats to its PISA 2015 data (OECD, 2016a; OECD, 2017b); among the replication candidates, its instructional items also carry the heaviest nonresponse burden outside Thailand, with weighted item nonresponse of 11.79% on IBTEACH, 11.76% on teacher support and 11.88% on teacher-directed instruction, against 0.10-0.33% in Viet Nam (Table S3). Malaysia was therefore retained as a sensitivity case rather than promoted to a primary comparison system.

Analytic samples and missingness. The inquiry models included students with valid science plausible values, final and replicate weights, IBTEACH, sex and ESCS. The primary Thailand complete-case sample comprised 7,250 students, representing an estimated weighted population of 545,378. Replication sample sizes were 5,687 in Singapore, 6,344 in Indonesia and 5,818 in Viet Nam; the Malaysia sensitivity sample comprised 8,115 students. The trend analyses used all students with the outcome and required design variables in each country-cycle cell. Unweighted sample sizes ranged from 6,115 to 13,439 across the nine primary cells (Table 1). These are analysis samples, not simple random samples, and population estimates incorporate the final student weight. Complete-case restriction in Thailand excluded 999 of the 8,249 students with valid outcome and design data (12.1% unweighted), driven primarily by nonresponse on the IBTEACH index (weighted item nonresponse 14.07%; Table S3, Figure S1), with ESCS nonresponse (2.21%) accounting for most of the remainder; teacher support (14.10%) and teacher-directed instruction (14.34%) were not required for the primary sample and were missing only for the moderator-specific models, which consequently

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

have smaller samples (n = 7,220 and n = 7,205; Table 3); the weighted complete-case response rate was 85.91%. The archived response-propensity diagnostics indicate that inclusion was associated with observed characteristics—female students (odds ratio = 1.58) and students one standard deviation higher in ESCS (odds ratio = 1.18) were modestly more likely to be retained—so the exclusions were not strictly random. The prespecified stabilised inverse response-propensity weights, clipped at 0.878 and 2.337 (respondent-weighted percentiles 1 and 99), were constructed to address precisely these observed differences, and the nonlinearity finding was unchanged under this reweighting ($\chi^2(2) = 33.95$, $p < .001$). Bias from unobserved nonresponse cannot be excluded by either analysis.

Measures

Science achievement. Science achievement was represented by all ten PISA science plausible values (PV1SCIE-PV10SCIE). Plausible values are multiple draws from an estimated proficiency distribution and must be analysed and combined rather than averaged once and treated as an error-free score.

Inquiry-oriented science instruction (IBTEACH). Inquiry-oriented science instruction was measured by the PISA 2015 IBTEACH index, a student-reported summary of how frequently nine inquiry-related practices occurred in science lessons (Table S9); higher values indicate more frequent exposure (OECD, 2016b). The nine items (stem ST098) are: ST098Q01TA ('Students are given opportunities to explain their ideas'), ST098Q02TA ('Students spend time in the laboratory doing practical experiments'), ST098Q03NA ('Students are required to argue about science questions'), ST098Q05TA ('Students are asked to draw conclusions from an experiment they have conducted'), ST098Q06TA ('The teacher explains how a <school science> idea can be applied to a number of different phenomena'), ST098Q07TA ('Students are allowed to design their own experiments'), ST098Q08TA ('There is a class debate about investigations'), ST098Q09TA ('The teacher clearly explains the relevance of <broad science> concepts to our lives') and ST098Q10NA ('Students are asked to do an investigation to test ideas'). Two of the nine items describe teacher explanation rather than student activity (ST098Q06TA and ST098Q09TA). The index captures reported frequency, not directly observed instructional quality, sequencing or fidelity. Its scale is therefore suitable for modelling graded associations but not for assigning a universally meaningful classroom dosage.

Moderators and adjustment variables. The four prespecified moderators were science self-efficacy (SCIEEFF), perceived teacher support in science (TEACHSUP), teacher-directed science instruction (TDTEACH) and the PISA index of economic, social and cultural status (ESCS). Higher values reflect more of the named construct; in particular, TDTEACH (stem ST103) summarises four teacher-led practices—explaining scientific ideas (ST103Q01NA), holding whole-class discussions (ST103Q03NA), discussing students' questions (ST103Q08NA) and demonstrating an idea (ST103Q11NA)—while TEACHSUP (stem ST100) summarises five perceived teacher-support items (ST100Q01TA-ST100Q05TA; Table S9). ESCS and a harmonised female indicator were the minimal adjustment set for RQ1-RQ3. Across all three cycles, the source metadata coded ST004D01T = 1 as female and 2 as male; nonresponse and invalid categories were excluded rather than recoded. For RQ4, the primary socioeconomic gradient used the protocol-authorised trend-compatible ESCS variables for 2015 and 2018 and the linked 2022 anchor; native-cycle ESCS gradients were retained as a sensitivity check.

Survey estimation and uncertainty

Each weighted least-squares model was fitted separately for all ten science plausible values using the final student weight. Sampling covariance was computed with all 80 balanced repeated replication weights and the PISA Fay coefficient of 0.5: four-eightieths of the sum of squared replicate deviations. The pooled covariance was the mean sampling covariance plus 1.1 times the between-plausible-value covariance. This procedure carries both complex-sample and achievement-imputation uncertainty into coefficients, predictions, contrasts and joint tests (OECD, 2017b; Rutkowski et al., 2010). We report two-sided 95% confidence intervals. P-values are evidential summaries rather than dichotomous proof, and substantive interpretation considers effect pattern, precision and sensitivity.

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

Design-based variance estimation rather than multilevel decomposition.The 80 BRR-Fay replicate weights are constructed by the OECD from PISA's actual stratified two-stage sample design, with schools as primary sampling units. Re-estimating every model across the full replicate set therefore carries the design effects of clustering and stratification into the sampling covariance of each coefficient, prediction and contrast. This design-based approach is the OECD-recommended standard for estimating population-averaged quantities from PISA (OECD, 2017b; Rutkowski et al., 2010), and the present research questions concern precisely such population-averaged marginal associations rather than the partitioning of achievement variance between schools and students. A hierarchical linear model with school random effects would answer a different question—how much achievement variance lies between schools, and how school-centred associations differ—and is not required for the estimands reported here. The residual possibility that school-level processes beyond those absorbed by the replicate weights shape the observed associations is acknowledged in the limitations.

RQ1: functional form and sensitivity

For Thailand, the frozen model sequence compared a linear term, a quadratic expansion and a restricted cubic spline (RCS). The primary RCS used four knots placed at weighted IBTEACH percentiles 5, 35, 65 and 95: -1.0923, -0.1210, 0.2956 and 1.5360. This placement follows Harrell's (2015) recommended defaults for four-knot restricted cubic splines, which sit at the 5th, 35th, 65th and 95th percentiles of the predictor distribution; locating knots at distributional quantiles rather than arbitrary values stabilises the spline in the sparse tails while retaining flexibility where observations are dense. The knot percentiles were prespecified, and the resulting knot values were derived from the weighted Thai IBTEACH distribution alone, without reference to the outcome. Sex and ESCS were included in each specification. Nonlinearity was evaluated by a joint Wald test of the nonlinear spline terms, not by selecting a turning point after inspecting the outcome. Score-point contrasts were prespecified at the weighted 5th, 50th and 95th IBTEACH percentiles so that practical magnitudes would not depend on the location of a fitted turning point, and are reported with design-based 95% confidence intervals under all three specifications (Table S11). Primary predictions were displayed over weighted percentiles 1-99. A P5-P95 trimmed analysis assessed tail influence. A second sensitivity used stabilised inverse response-propensity weights estimated from observed covariates and clipped at respondent-weighted percentiles 1 and 99. Neither sensitivity can eliminate bias from unobserved nonresponse.

RQ2: moderation and multiplicity

Each moderator was evaluated in a separate RCS model by interacting the moderator with the three inquiry basis terms. The prespecified three-degree-of-freedom joint interaction test asked whether the entire fitted inquiry function differed across moderator values. Four p-values were adjusted with the Benjamini-Hochberg false-discovery-rate procedure (Benjamini & Hochberg, 1995). Predictions at the weighted 25th, 50th and 75th percentiles were used to interpret supported interactions. Because spline interactions can change curvature and slope simultaneously, the joint test and plotted uncertainty bands are more informative than any isolated product-term coefficient.

RQ3: replication and common support

The Thailand RCS specification was refitted separately in Singapore, Indonesia and Viet Nam, with Malaysia added only as a sensitivity case. Country curves were plotted and compared over the intersection of their empirical IBTEACH support, -0.8385 to 1.1777. This restriction reduces extrapolation but does not make countries exchangeable. We therefore treated replication as evidence about the presence and form of nonlinearity within each system, not as a test that country intercepts or curves should coincide.

RQ4: repeated cross-sectional context

Science means were estimated by country and cycle using all ten plausible values and the full survey design. Within-country contrasts were 2018 minus 2015, 2022 minus 2018 and 2022 minus 2015. Because samples were independent, contrast variances were sums of the two cycle-specific variances. Socioeconomic gradients were coefficients for trend-compatible ESCS in sex-adjusted

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

models, interpreted as the expected PISA science-point difference per one-unit ESCS difference. Difference-in-change contrasts compared Thailand with Singapore and Indonesia. With only three cross-sections, we did not fit a continuous time trajectory and did not attribute the 2022 contrast to a specific historical event.

Preregistration status, data integrity, and reproducibility

Throughout this article, 'prespecified' refers to an internal rather than a public lock. The analysis plan was internally locked (Gate 0G) prior to executing the code, but it was not publicly preregistered on a platform like OSF prior to data access. To ensure full transparency, the locked SAP, code, and outputs are permanently archived at https://zenodo.org/records/23853258.

Seven prespecified readiness gates were passed before substantive interpretation: authoritative-file selection, required-variable availability, country-cycle eligibility, construct harmonisation, official benchmark reconciliation, empirical support and statistical-analysis-plan freezing (Table S1). All nine primary country-cycle science means and standard errors matched published OECD values within locked rounding tolerances (Table S2). Every primary cell contained ten plausible values and 80 replicate weights, and no duplicate student rows were detected. The executed notebook contained no errors; prediction supports, confidence intervals and p-value bounds were programmatically checked. The analysis code, output tables, notebook, and validation report form the reproducibility archive, which is permanently deposited at https://zenodo.org/records/23853258.

Table 1: Analytic systems, cycles, and unweighted sample sizes

System	Cycle	Role	Students	Schools
Thailand	2015	Focal	8,249	273
Thailand	2018	Focal	8,633	290
Thailand	2022	Focal	8,495	279
Singapore	2015	Comparison	6,115	177
Singapore	2018	Comparison	6,676	166
Singapore	2022	Comparison	6,606	164
Indonesia	2015	Comparison	6,513	236
Indonesia	2018	Comparison	12,098	397
Indonesia	2022	Comparison	13,439	410
Viet Nam	2015	Replication only	5,818	-
Malaysia	2015	Sensitivity only	8,115	-

Note.RQ1-RQ3 analytic samples require complete instructional and adjustment variables and therefore differ from the all-outcome RQ4 samples, for which school counts are shown.

RESULTS AND DISCUSSION

RQ1: Thailand's inquiry-achievement function

A linear specification explained 10.01% of weighted variation in Thailand's complete-case sample. Both more flexible specifications improved the descriptive fit: the quadratic model yielded weighted $R^2 = .106$ and a nonlinear Wald $\chi^2(1) = 31.83$, $p < .001$, while the primary RCS yielded weighted $R^2 = .108$ and $\chi^2(2) = 37.57$, $p < .001$ (Table 2; Table S4). The fitted curve increased from low inquiry values into the middle-positive range, flattened and then changed direction at higher reported frequencies (Figure 1). Its descriptive grid maximum occurred near IBTEACH = 0.41. That location is not interpreted as a classroom target because tail uncertainty, scaling and residual confounding prevent a policy-dosage claim.

Nonlinearity survived both prespecified sensitivities. Restricting the sample to P5-P95 produced $\chi^2(2) = 18.80$, $p < .001$ (n = 6,538), and response-propensity weighting produced $\chi^2(2) = 33.95$, $p <$

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

.001 (n = 7,250). The fitted high points were broadly similar (approximately 0.44 in the trimmed analysis and 0.33 under response weighting), but their differences reinforce the decision to emphasise the overall shape rather than a single turning value. Confidence intervals widened toward the distributional tails.

In practical score-point terms, predicted achievement at the fitted peak (IBTEACH ≈ 0.41, between the weighted 50th and 75th percentiles of the Thai distribution) exceeded the prediction at the weighted 95th percentile (IBTEACH = 1.536) by 13.65 points (95% CI [9.18, 18.11]; 431.6 vs 418.0 points on the PISA science scale); by the 99th percentile the fitted gap reached 42.96 points (95% CI [29.76, 56.17]). Because the peak was located after the model was fitted, these magnitudes—read from the archived prediction grid (Additional File 2)—are descriptive summaries of the fitted curve rather than inferential contrasts or causal returns to instruction (Table S11). The trimmed curve in Figure 1 differs from the primary curve within the retained P5-P95 range and not only in the tails: at the 95th-percentile knot it is about 20 points lower and at the 5th-percentile knot about 12 points higher (Table S13). The knots were not recomputed for the trimmed model; they were held at the full-sample values (-1.0923, -0.1210, 0.2956, 1.5360) so that the two specifications would remain directly comparable. Refitting the trimmed subsample with knots taken from its own distribution (-0.7004, -0.1210, 0.2661, 1.0466) changed the curve very little, giving predicted levels of 434.7, 429.1 and 401.3 points at the three anchors against 435.4, 429.1 and 398.0 under fixed knots, and left nonlinearity intact, chi2(2) = 21.16, p < .001. The ~20-point divergence at P95 therefore reflects primarily the removal of the 712 tail students from the fit (~16.7 points) alongside a smaller boundary-knot shift (~3.3 points) when outer knots sit at the truncation edges without exterior tail observations (Harrell, 2015). The trimmed specification is correspondingly unstable at the low end, where its P50 minus P5 contrast is -6.26 points (95% CI [-19.89, 7.38]) against +6.66 points in the complete-case fit, and it is reported as a sensitivity check rather than as an alternative estimate of the curve.

Because the peak-based gaps depend on a turning point selected after fitting, prespecified contrasts were also computed at the weighted 5th, 50th and 95th IBTEACH percentiles, which were fixed before the models were run. Predicted achievement was 6.66 points higher at the median than at the 5th percentile (95% CI [2.54, 10.79], p = .002) and 11.65 points higher at the median than at the 95th percentile (95% CI [5.81, 17.50], p < .001); the direct 5th-versus-95th-percentile contrast was smaller and not clearly resolved (4.99 points, 95% CI [-2.51, 12.49], p = .192). Both prespecified contrasts were reproduced under response-propensity weighting (+6.15 points, 95% CI [1.87, 10.43]; and +12.33 points, 95% CI [6.31, 18.36]), and the upper-end contrast was reproduced with a larger magnitude under trimmed support (+31.07 points, 95% CI [15.09, 47.05]; Table S11). Non-monotonicity therefore does not depend on locating a peak.

Table 2: Thailand PISA 2015 functional-form models

Model	Specification	n	Weighted R^2	Nonlinear χ^2	df	p
Linear	Complete case	7,250	0.100	-	-	-
Quadratic	Complete case	7,250	0.106	31.83	1	< .001
Restricted cubic spline	Complete case	7,250	0.108	37.57	2	< .001
Restricted cubic spline	P5-P95 trimmed	6,538	0.104	18.80	2	< .001
Restricted cubic spline	Response propensity	7,250	0.102	33.95	2	< .001

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

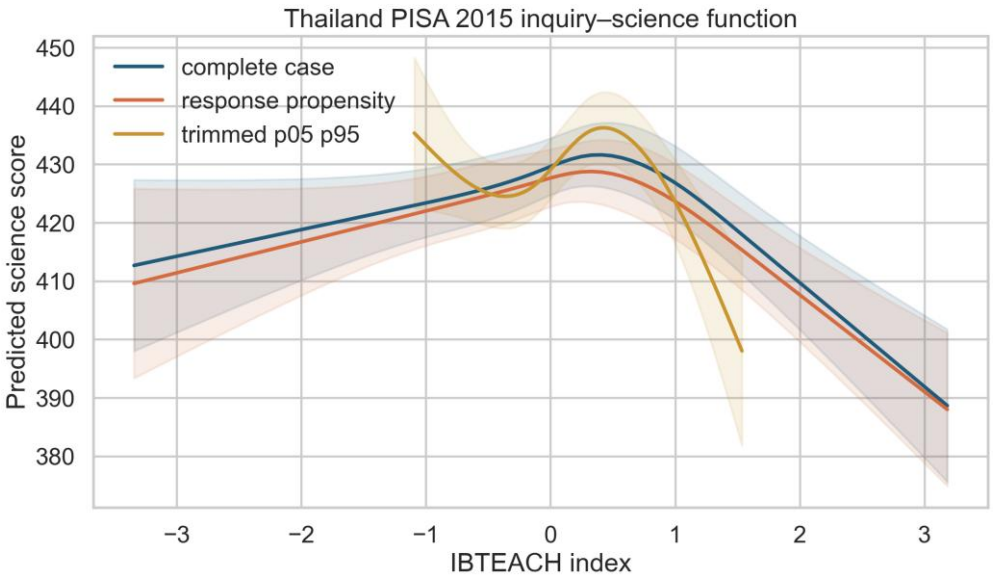


Figure 1: Adjusted association between inquiry-based science instruction (IBTEACH) and science achievement in Thailand, PISA 2015. Curves show the primary restricted cubic spline and the prespecified P5-P95 and response-propensity sensitivities with 95% confidence intervals. Tail maxima are descriptive fitted features, not instructional targets. The trimmed curve is fitted on the P5-P95 subsample with knots held at full-sample values, so it diverges from the primary curve within the retained range as well as in the tails (see text); the divergence reflects the excluded tail observations rather than a behavioural shift.

RQ2: Conditionality

After correction across four prespecified tests, the inquiry function varied with teacher-directed instruction, $\chi^2(3) = 48.46$, false-discovery-rate-adjusted $p < .001$; teacher support, $\chi^2(3) = 11.98$, adjusted $p = .015$; and ESCS, $\chi^2(3) = 8.96$, adjusted $p = .040$ (Table 3; Table S5; Figure 2). Science self-efficacy did not moderate the function, $\chi^2(3) = 0.85$, adjusted $p = .838$. The clearest pattern appeared for teacher-directed instruction: at high inquiry values, fitted achievement was less depressed where teacher direction was higher, although uncertainty increased at the extremes. Teacher support showed a similar but weaker buffering pattern. ESCS altered both score level and curvature; it did not remove the downturn at high inquiry frequency. These comparisons describe conditional associations, not evidence that increasing teacher direction or support would necessarily change the curve for the same students.

One feature of the measures qualifies that reading. IBTEACH is not composed only of student-led activity: two of its nine items ask how often the teacher explains how a science idea can be applied and how often the teacher explains the relevance of broad science concepts (Table S9). IBTEACH and TDTEACH are drawn from different questionnaire blocks and share no items, and their weighted correlation in Thailand is modest ($r = 0.227$, 95% CI [0.185, 0.269]), as is the IBTEACH-teacher support correlation ($r = 0.326$, 95% CI [0.297, 0.355]); the correlations are of similar size in the replication systems (Table S10). The two constructs therefore share only about 5% of their variance, so the teacher-directed interaction is not an artefact of overlapping item content. At the same time, part of what the index counts as inquiry is teacher explanation, which sits alongside Cairns's (2019) item-level result that teacher-explanation items were among the most positively associated with achievement. The moderation finding is accordingly read as evidence that the inquiry-achievement function varies with the amount of teacher direction students report, not as a contrast between student-led and teacher-led instruction.

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

Table 3: Joint tests of moderation of the Thailand inquiry function

Moderator	n	χ^2	df	Raw p	BH-FDR p
Science self-efficacy	7,168	0.85	3	= .838	= .838
Teacher support	7,220	11.98	3	= .007	= .015
Teacher-directed instruction	7,205	48.46	3	< .001	< .001
Socioeconomic status (ESCS)	7,250	8.96	3	= .030	= .040

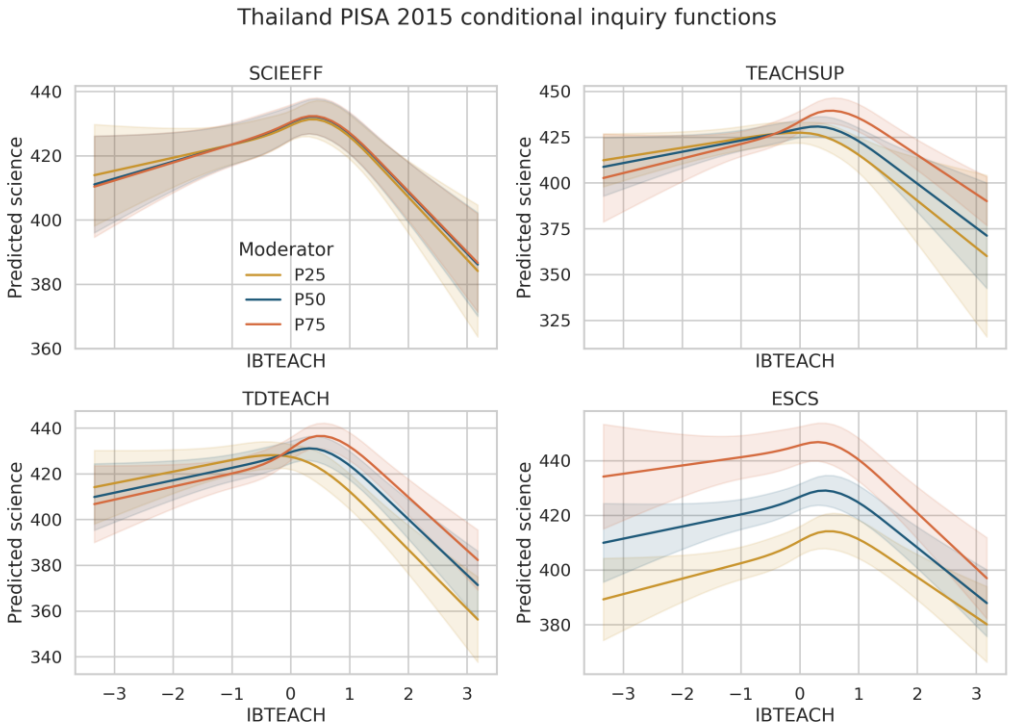


Figure 2: Thailand's adjusted inquiry-achievement function at the weighted 25th, 50th, and 75th percentiles of science self-efficacy, teacher support, teacher-directed instruction, and ESCS. Shaded bands are 95% confidence intervals. Joint tests use three degrees of freedom and Benjamini-Hochberg correction across four moderators.

RQ3: Cross-system replication

Evidence of nonlinearity was strong in Thailand, $\chi^2(2) = 37.57$, $p < .001$; Singapore, $\chi^2(2) = 149.45$, $p < .001$; and Indonesia, $\chi^2(2) = 49.82$, $p < .001$ (Table 4). Viet Nam's joint test was not clear at the conventional .05 level, $\chi^2(2) = 5.01$, $p = .082$. This result is best read as inconclusive rather than as evidence of linearity: Viet Nam's analytic sample (n = 5,818) is comparable to the other systems, but its weighted IBTEACH distribution is markedly more compressed (1st-99th percentiles from -1.66 to 1.90, versus -3.34 to 3.18 in Thailand) and curvature is harder to detect over a narrower effective range in a single cross-section; the common-support restriction applies only to the display in Figure 3, whereas the Wald test uses each system's full fitted model and is unaffected by it. As a sensitivity analysis, each replication system was refitted with knots placed at its own weighted IBTEACH percentiles 5, 35, 65 and 95 rather than at Thailand's. Every conclusion was unchanged: Singapore chi2(2) = 151.06, p < .001; Indonesia chi2(2) = 52.38, p < .001; Malaysia chi2(2) = 24.83, p < .001; and Viet Nam chi2(2) = 4.64, p = .098 (Table S12). Viet Nam's result therefore remains inconclusive under its own knot placement, and the replication pattern does not depend on importing Thailand's knot locations; sampling variance alone could plausibly produce the observed p = .082. Malaysia, retained as sensitivity-only, also showed nonlinearity, $\chi^2(2) = 28.49$, $p < .001$. Over shared support, the curves differed in level and shape (Figure 3). The replicated conclusion is therefore bounded: a single linear slope is inadequate in several systems, but the data do not support one regional threshold or an interchangeable instructional response function.

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

Table 4: Country-specific restricted cubic spline models, PISA 2015

System	Role	n	Weighted R^2	Nonlinear χ^2	df	p
Thailand	Focal	7,250	0.108	37.57	2	< .001
Singapore	Replication	5,687	0.183	149.45	2	< .001
Indonesia	Replication	6,344	0.161	49.82	2	< .001
Viet Nam	Replication	5,818	0.111	5.01	2	= .082
Malaysia	Sensitivity	8,115	0.129	28.49	2	< .001

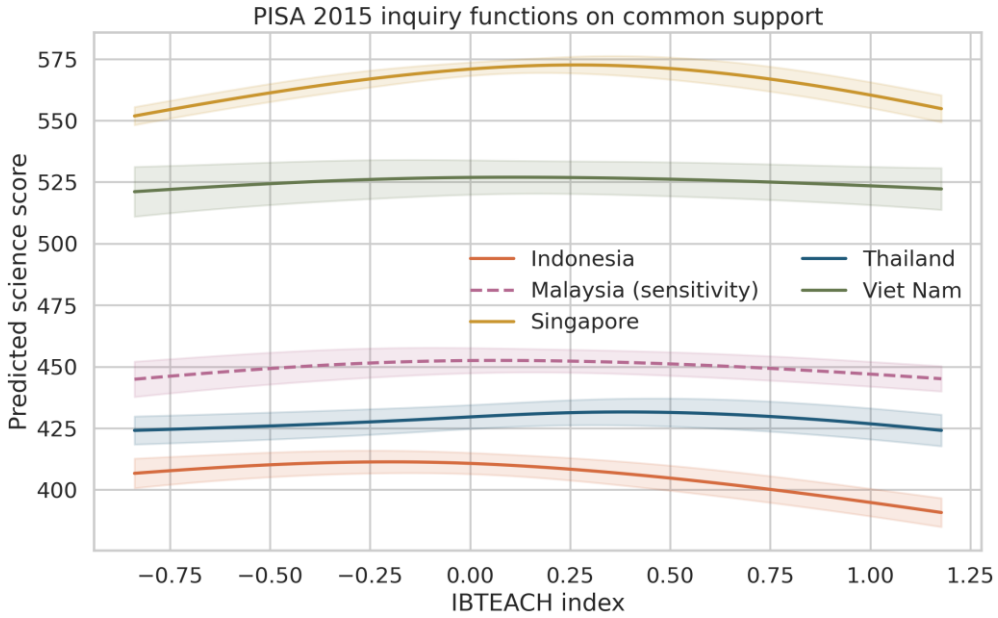


Figure 3: Country-specific adjusted inquiry-achievement curves in Thailand, Singapore, Indonesia, and Viet Nam, with Malaysia shown as a sensitivity case, PISA 2015. Curves are displayed only over common IBTEACH support [-0.8385, 1.1777]; bands are shaded 95% confidence intervals.

RQ4: Changing achievement and socioeconomic gradients

Thailand's estimated science mean was 421.34 (SE = 2.83) in 2015, 425.81 (SE = 3.18) in 2018 and 409.26 (SE = 2.78) in 2022. The 2018-2015 difference was +4.48 points, 95% CI [-3.87, 12.82], p = .293. The 2022 mean was 16.55 points lower than 2018, 95% CI [-24.83, -8.27], p < .001, and 12.07 points lower than 2015, 95% CI [-19.86, -4.29], p = .002 (Table 5; Table S6; Figure 4). These are population-level differences between independent samples, not within-student losses.

The comparison systems followed different paths. Singapore declined by 4.64 points from 2015 to 2018, 95% CI [-8.37, -0.90], then increased by 10.50 points from 2018 to 2022, 95% CI [6.59, 14.40], finishing 5.86 points above 2015. Indonesia declined by 7.03 points from 2015 to 2018, 95% CI [-13.90, -0.16], and by a further 13.21 points from 2018 to 2022, 95% CI [-20.07, -6.35]. Across 2015-2022, Indonesia's mean decreased by 20.24 points, 95% CI [-27.34, -13.13], p < .001. Thailand's 2022-2018 change differed from Singapore's by -27.05 points, 95% CI [-36.20, -17.89], but did not clearly differ from Indonesia's.

Thailand's trend-compatible ESCS gradient was 21.34 science points in 2015, 24.64 in 2018 and 23.06 in 2022. Its 2015-2022 change was +1.72 points per ESCS unit, 95% CI [-4.27, 7.71], p = .574 (Table 6; Table S7). Singapore's gradient rose by 8.79 between 2018 and 2022, 95% CI [4.08, 13.49], while Indonesia's 2022 gradient was 8.44 points lower than in 2015, 95% CI [-13.24, -3.65]. Across the full 2015-2022 window, Thailand's net gradient change was statistically compatible with Singapore's (+4.37, 95% CI [-0.30, 9.04], p = .067), whereas Indonesia's gradient contracted by a clear 8.44 points (95% CI [-13.24, -3.65], p < .001). The Thailand-Indonesia difference in 2015-2022 gradient change was +10.16, 95% CI [2.49, 17.84]. A smaller gradient does not by itself indicate an

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

equity improvement when mean performance is also falling; the level and gradient results must be interpreted jointly.

Table 5: Survey-weighted science means by system and cycle

System	Cycle	n	Mean	SE	95% CI
Thailand	2015	8,249	421.34	2.83	[415.78, 426.89]
Thailand	2018	8,633	425.81	3.18	[419.58, 432.04]
Thailand	2022	8,495	409.26	2.78	[403.81, 414.72]
Singapore	2015	6,115	555.57	1.20	[553.23, 557.92]
Singapore	2018	6,676	550.94	1.48	[548.03, 553.84]
Singapore	2022	6,606	561.43	1.33	[558.83, 564.03]
Indonesia	2015	6,513	403.10	2.57	[398.07, 408.13]
Indonesia	2018	12,098	396.07	2.39	[391.39, 400.75]
Indonesia	2022	13,439	382.86	2.56	[377.84, 387.88]

Table 6: Trend-compatible socioeconomic gradients in science achievement

System	Cycle	n	Gradient	SE	95% CI	Net 2015-2022 change
Thailand	2015	8,090	21.34	2.17	[17.08, 25.59]	+1.72, 95% CI [-4.27, 7.71], p = .574
Thailand	2018	8,582	24.64	1.97	[20.77, 28.50]	—
Thailand	2022	8,436	23.06	2.15	[18.84, 27.27]	—
Singapore	2015	6,089	44.72	1.49	[41.79, 47.64]	+4.37, 95% CI [-0.30, 9.04], p = .067
Singapore	2018	6,647	40.30	1.52	[37.32, 43.27]	—
Singapore	2022	6,559	49.08	1.86	[45.44, 52.73]	—
Indonesia	2015	6,507	21.84	1.70	[18.51, 25.17]	−8.44, 95% CI [-13.24, -3.65], p < .001
Indonesia	2018	12,008	16.72	2.07	[12.66, 20.78]	—
Indonesia	2022	13,303	13.39	1.76	[9.95, 16.84]	—

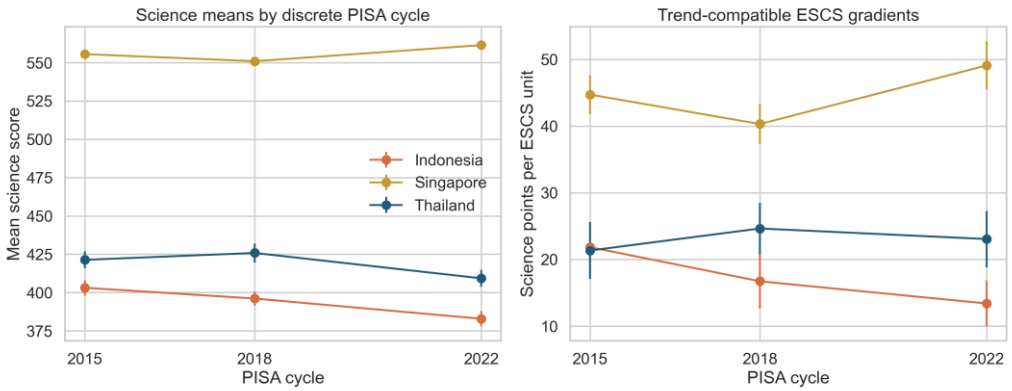


Figure 4: Survey-weighted mean science achievement and sex-adjusted, trend-compatible ESCS gradients in Thailand, Singapore, and Indonesia, PISA 2015, 2018, and 2022. Error bars are 95% confidence intervals. Cycle samples are independent cross-sections.

This study asked when inquiry is associated with science achievement rather than whether inquiry is globally ‘good’ or ‘bad’. Three findings answer that question. First, Thailand's 2015

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

association was robustly nonlinear under the prespecified spline, trimmed-support and response-propensity analyses. Second, the shape depended on teacher-directed instruction, teacher support and socioeconomic status, but not on science self-efficacy after multiplicity adjustment. Third, nonlinearity replicated in Singapore and Indonesia but not clearly in Viet Nam, and the common-support curves did not converge on a universal regional threshold. The broader 2015-2022 context also diverged: Thailand's 2022 science mean was below both earlier cycles without a clear change in its socioeconomic gradient, Singapore recovered above its 2015 mean, and Indonesia's mean and gradient both decreased.

A conditional, guidance-sensitive account

The nonlinear Thailand result is consistent with the proposition that inquiry frequency and inquiry quality are not interchangeable. At low reported exposure, additional opportunities to interpret evidence and explain ideas may coincide with richer engagement. At high reported frequency, the same index may represent very different classroom realities: well-scaffolded investigation, repetitive activity, insufficient consolidation or student perceptions shaped by difficulty. The possibility that exceptionally high IBTEACH scores partly reflect perceived task difficulty or unguided activity rather than effective inquiry is treated as an explicit construct-validity limitation below. A flexible curve can reveal that the marginal association changes across the range, but it cannot adjudicate among these mechanisms. This is why the turning region should be treated as a descriptive feature of this sample rather than a recommended frequency.

The moderation evidence sharpens that interpretation. At high inquiry values, higher teacher-directed instruction and teacher support were associated with a less adverse fitted pattern. The result fits theoretical and meta-analytic work in which guidance makes inquiry more productive by structuring search, connecting observations to concepts and providing timely feedback (Hmelo-Silver et al., 2007; Lazonder & Harmsen, 2016). It also aligns with PISA studies separating guided and independent inquiry (Aditomo & Klieme, 2020). The finding should not be simplified into a contest between teacher-led and student-led instruction. The supported interaction suggests complementarity: students may exercise epistemic agency within deliberately organised instruction.

ESCS moderation points to unequal instructional conditions or unequal returns to the same reported frequency. Students' prior opportunities, language resources, material support and familiarity with school-valued forms of reasoning may influence how readily they benefit from open-ended science activity. Yet the interaction does not reveal which resource matters, and the PISA ESCS index is not a manipulable treatment; evidence on whether inquiry-based teaching narrows or widens socioeconomic achievement gaps remains mixed (Tang et al., 2019), so the present result should not be read as resolving that question. The non-supported self-efficacy interaction is also informative. Although self-efficacy was theoretically plausible, the data did not show that it systematically altered the entire spline after false-discovery-rate correction. The false-discovery-rate procedure controls the expected proportion of false positives among the four prespecified tests (Benjamini & Hochberg, 1995); selective rather than universal moderation under that correction reduces the likelihood that the supported interactions are artefacts of multiple testing.

Why transportability must be tested

As argued in the Introduction, a shape estimated in one system cannot be assumed to define a regional threshold; transportability is an empirical question, and the cross-system results bear directly on it. Thailand, Singapore and Indonesia each rejected a purely linear description, but they differed in score level, curvature and support. Viet Nam did not provide clear evidence of nonlinearity under the same test. Such variation is compatible with differences in curriculum, classroom routines, index interpretation and student composition, as well as sampling uncertainty. Prior work has shown that the PISA inquiry construct can decompose differently across systems and that disaggregated practices have heterogeneous associations (Aditomo & Klieme, 2020; Cairns, 2019; Lau & Lam, 2017). The appropriate policy question therefore remains the one posed in the Introduction—not 'How much inquiry does Southeast Asia need?' but 'Which forms of inquiry, with what guidance, for which learners, under which system conditions?'

Because visual comparisons outside overlapping IBTEACH ranges would be driven by extrapolation, the common-support restriction was prespecified. Even within common support, however, country contrasts remain observational. System-specific differences can reflect

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

unmeasured school composition, language and response styles, opportunity to learn and the organisation of science courses. Replication here means that the need for flexible functional form recurs; it does not mean that the same intervention would produce the observed curves.

Reading the cross-cycle context without causal overreach

Thailand's lower 2022 mean provides a consequential context for science education, but it does not validate a causal explanation based on 2015 inquiry practices. The instructional and temporal layers involve different comparisons. Layer A relates students' reported 2015 classroom experience to contemporaneous achievement. Layer B compares independent national cross-sections across three cycles. It cannot connect particular 2015 students, teachers or schools to the 2022 population. Similarly, labelling the 2022 difference a pandemic effect would require a counterfactual design and stronger assumptions than these data permit.

The joint reading of means and socioeconomic gradients is more informative than either metric alone. Thailand's gradient remained statistically compatible across 2015-2022 while the mean fell in 2022, suggesting that the performance shift was not accompanied by a clearly estimable change in the linear ESCS association. Indonesia's gradient decreased alongside a substantial mean decline, so the smaller gradient should not automatically be framed as equitable progress. Singapore's post-2018 mean increase coincided with a steeper gradient. These contrasting configurations demonstrate why equity requires attention to both distributional level and social patterning.

Implications for science education

For practice, the findings are consistent with designed inquiry over frequency targets. Instructional planning should specify the scientific practice, the conceptual knowledge learners need, the scaffolds available, how evidence will be discussed and when teachers will consolidate emerging ideas. In prior instructional-design research, frameworks such as the 5E instructional model (Bybee et al., 2006) and Claim-Evidence-Reasoning (CER) scaffolding (McNeill & Krajcik, 2012) illustrate how teacher explanation, modelling and feedback can coexist with student questioning and investigation. Monitoring systems should therefore avoid using a single inquiry-frequency index as a quality score. Classroom observation, task analysis and teacher reports are needed to distinguish productive guidance from either procedural scripting or unsupported discovery.

For research, the next step is to test mechanisms directly. Longitudinal or experimental studies could manipulate forms and timing of guidance while measuring prior knowledge, reasoning processes and implementation. Cross-national measurement work should examine whether inquiry items function comparably before comparing latent levels. Where large-scale surveys remain the data source, analysts should use plausible values and replicate weights, model nonlinearity before interpreting average slopes, restrict comparisons to common support and publish prediction-level uncertainty. Mixed-method studies in Thailand and the comparison systems could investigate what students mean when they report frequent inquiry and how those reports map onto observed classroom practice.

For policy, the strongest defensible conclusion is conditional. Inquiry-oriented science education should not be abandoned because a linear PISA coefficient is negative, nor promoted through a universal numerical benchmark (Sjøberg, 2018). Systems can instead invest in teachers' capacity to sequence investigation, explanation and reflection; provide resources that allow all students to participate meaningfully; and evaluate implementation with outcomes broader than one cross-sectional test score. The present results identify where more targeted evidence is needed, not a one-size-fits-all prescription.

Strengths and limitations

The study's strengths include a protocol-defined two-layer design, a flexible functional-form analysis, prespecified moderators with multiplicity control, common-support replication, all ten plausible values, the full BRR-Fay design and benchmark reconciliation against OECD publications. Sensitivity analyses addressed tail influence and observed response differences. Separating the inquiry and trend layers also prevents an otherwise tempting ecological causal claim.

Several limitations remain. First, PISA is observational and cross-sectional; reciprocal relationships, selection and omitted variables can shape associations. Second, IBTEACH is a student-reported frequency index that cannot fully distinguish guidance, quality or content. Cross-system

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

measurement equivalence was not established in this analysis. Construct validity at the upper end of the scale deserves explicit caution: exceptionally high reported inquiry frequency may reflect student perceptions of task difficulty, unguided discovery activity or response styles rather than effective inquiry, so self-report bias may contribute to the high-frequency downturn. Third, complete cases are primary, and response-propensity weighting addresses only observed response differences. Fourth, the survey design was handled through weights and replicates, but school random effects were not separately modelled. Fifth, spline tails are imprecise, and fitted maxima are sensitive descriptive summaries. Sixth, cross-cycle score and ESCS comparisons can be affected by changes in population composition and measurement context. Seventh, only Thailand's inquiry model was built from its own distributional criteria; the comparison systems replicate the Thai specification—including the Thai-derived knot percentiles—so country curves are not independently derived, and replication strengthens comparability at the cost of system-specific functional-form exploration. A sensitivity analysis that refits each system with its own percentile knots leaves every conclusion unchanged (Table S12), although the primary country models remain specification-transported. Three cycles do not constitute individual trajectories or a causal interrupted time series. Finally, the selected systems provide analytically useful contrasts but do not represent all of Southeast Asia.

CONCLUSION AND IMPLICATIONS

Thailand's PISA 2015 data show a robust nonlinear association between reported inquiry-oriented science instruction and achievement. Its shape varies with teacher direction, teacher support and socioeconomic status, and nonlinearity is evident in some-but not all-comparison systems. The result is not a universal threshold. It supports a more precise proposition: the results are consistent with the proposition that inquiry's association with achievement depends heavily on its instructional configuration, including how guidance and learner resources are aligned with demanding scientific practices. The changing 2015-2022 achievement context strengthens the urgency of that question but does not identify a causal link from earlier instruction to later national performance. For Southeast Asian science educators, the practical message is one this study is consistent with rather than one it establishes: prior guidance research supports deliberately configured inquiry, in which investigations are preceded by the conceptual knowledge they require, scaffolded during data collection and interpretation, and consolidated through teacher explanation and feedback (Bybee et al., 2006; Furtak et al., 2012; Hmelo-Silver et al., 2007; Lazonder & Harmsen, 2016; McNeill & Krajcik, 2012). Sequencing, scaffolding and consolidation were not measured in these data, so the present results cannot estimate what they would add. Frequency alone is a poor proxy for quality, and monitoring frameworks that reward reported frequency risk misclassifying unsupported activity as effective inquiry. Carefully designed classroom studies are needed to turn the present conditional associations into actionable evidence.

ACKNOWLEDGEMENTS

The author acknowledges the Organisation for Economic Co-operation and Development for providing public-use PISA data and technical documentation.

ETHICAL CONSIDERATIONS

This secondary analysis used anonymised, publicly available PISA public-use data. It involved no new participant recruitment, direct participant contact, intervention, or access to identifiable information. Formal human-participant ethics approval and informed consent were therefore not applicable to this analysis.

DATA AND CODE AVAILABILITY

PISA 2015, 2018, and 2022 public-use data are available from the OECD PISA database (https://www.oecd.org/en/about/programmes/pisa/pisa-data.html). The analysis code, derived tables, notebook, and validation reports accompany this submission both as supplementary files and at https://zenodo.org/records/23853258 ; raw OECD microdata are not redistributed in the archive. The

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

analyses added in revision (prespecified contrasts, construct correlations and system-specific knot refits) are deposited in the same archive.

USE OF ARTIFICIAL INTELLIGENCE TOOLS

OpenAI Codex using GPT-5.6 Sol supported Python code generation and execution, analytical workflow checks, drafting, language editing, and document formatting. It was not used to fabricate data or replace the author's intellectual interpretation. The author reviewed the analytical outputs and manuscript, verified the cited sources, and retains full responsibility for the accuracy, originality, and conclusions of the work. The author independently verified all statistical outputs, reviewed the theoretical framing, and takes full responsibility for the interpretation of the spline curves and cross-cycle contrasts. No artificial-intelligence system is listed as an author.

REFERENCES

Aditomo, A., & Klieme, E. (2020). Forms of inquiry-based science instruction and their relations with learning outcomes: Evidence from high and low-performing education systems. *International Journal of Science Education, 42*(4), 504-525. https://doi.org/10.1080/09500693.2020.1716093

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

Bybee, R. W., Taylor, J. A., Gardner, A., Van Scotter, P., Powell, J. C., Westbrook, A., & Landes, N. (2006). *The BSCS 5E instructional model: Origins and effectiveness.* BSCS.

Cairns, D. (2019). Investigating the relationship between instructional practices and science achievement in an inquiry-based learning environment. *International Journal of Science Education, 41*(15), 2113-2135. https://doi.org/10.1080/09500693.2019.1660927

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

Harrell, F. E., Jr. (2015). Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis (2nd ed.). Springer. https://doi.org/10.1007/978-3-319-19425-7

Hmelo-Silver, C. E., Duncan, R. G., & Chinn, C. A. (2007). Scaffolding and achievement in problem-based and inquiry learning: A response to Kirschner, Sweller, and Clark (2006). *Educational Psychologist, 42*(2), 99-107. https://doi.org/10.1080/00461520701263368

Hopfenbeck, T. N., Lenkeit, J., El Masri, Y., Cantrell, K., Ryan, J., & Baird, J.-A. (2018). Lessons learned from PISA: A systematic review of peer-reviewed articles on the Programme for International Student Assessment. *Scandinavian Journal of Educational Research, 62*(3), 333-353. https://doi.org/10.1080/00313831.2016.1258726

Jerrim, J., Oliver, M., & Sims, S. (2019). The relationship between inquiry-based teaching and students' achievement: New evidence from a longitudinal PISA study in England. *Learning and Instruction, 61*, 35-44. https://doi.org/10.1016/j.learninstruc.2018.12.004

Jiang, F., & McComas, W. F. (2015). The effects of inquiry teaching on student science achievement and attitudes: Evidence from propensity score analysis of PISA data. *International Journal of Science Education, 37*(3), 554-576. https://doi.org/10.1080/09500693.2014.1000426

Kirschner, P. A., Sweller, J., & Clark, R. E. (2006). Why minimal guidance during instruction does not work: An analysis of the failure of constructivist, discovery, problem-based, experiential, and inquiry-based teaching. *Educational Psychologist, 41*(2), 75-86. https://doi.org/10.1207/s15326985ep4102_1

Lau, K.-C., & Lam, T. Y.-P. (2017). Instructional practices and science performance of 10 top-performing regions in PISA 2015. *International Journal of Science Education, 39*(15), 2128-2149. https://doi.org/10.1080/09500693.2017.1387947

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.

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

McConney, A., Oliver, M. C., Woods-McConney, A., Schibeci, R., & Maor, D. (2014). Inquiry, engagement, and literacy in science: A retrospective, cross-national analysis using PISA 2006. *Science Education, 98*(6), 963-980. https://doi.org/10.1002/sce.21135

McNeill, K. L., & Krajcik, J. S. (2012). *Supporting Grade 5-8 students in constructing explanations in science: The claim, evidence, and reasoning framework for talk and writing.* Pearson.

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

OECD. (2016a). *PISA 2015 results (Volume I): Excellence and equity in education*. OECD Publishing. https://doi.org/10.1787/9789264266490-en

OECD. (2016b). *PISA 2015 results (Volume II): Policies and practices for successful schools*. OECD Publishing. https://doi.org/10.1787/9789264267510-en

OECD. (2017a). *PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving (revised edition)*. OECD Publishing. https://doi.org/10.1787/9789264281820-en

OECD. (2017b). *PISA 2015 technical report*. OECD Publishing. https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-report/PISA2015_TechRep_Final.pdf

OECD. (2019). *PISA 2018 results (Volume I): What students know and can do*. OECD Publishing. https://doi.org/10.1787/5f07c754-en

OECD. (2023). *PISA 2022 results (Volume I): The state of learning and equity in education*. OECD Publishing. https://doi.org/10.1787/53f23881-en

Oliver, M., McConney, A., & Woods-McConney, A. (2021). The efficacy of inquiry-based instruction in science: A comparative analysis of six countries using PISA 2015. Research in *Science Education, 51*(Suppl 2), 595-616. https://doi.org/10.1007/s11165-019-09901-0

Rutkowski, L., Gonzalez, E., Joncas, M., & von Davier, M. (2010). International large-scale assessment data: Issues in secondary analysis and reporting. *Educational Researcher, 39*(2), 142-151. https://doi.org/10.3102/0013189X10363170

Sjøberg, S. (2018). The power and paradoxes of PISA: Should inquiry-based science education be sacrificed to climb on the rankings? Nordic Studies in *Science Education, 14*(2), 186-202. https://doi.org/10.5617/nordina.6185

Tang, N.-E., Tsai, C.-L., Barrow, L., & Romine, W. (2019). Impacts of enquiry-based science teaching on achievement gap between high-and-low SES students: Findings from PISA 2015. *International Journal of Science Education, 41*(4), 448-470. https://doi.org/10.1080/09500693.2018.1555726

Teig, N., Scherer, R., & Nilsen, T. (2018). More isn't always better: The curvilinear relationship between inquiry-based teaching and student achievement in science. *Learning and Instruction, 56*, 20-29. https://doi.org/10.1016/j.learninstruc.2018.02.006

Teig, N., Scherer, R., & Olsen, R. V. (2022). A systematic review of studies investigating science teaching and learning: Over two decades of TIMSS and PISA. *International Journal of Science Education, 44*(12), 2035-2058. https://doi.org/10.1080/09500693.2022.2109075

von Davier, M., Gonzalez, E., & Mislevy, R. (2009). What are plausible values and why are they useful? *IERI Monograph Series, 2*, 9-36. https://www.ierinstitute.org/fileadmin/Documents/IERI_Monograph/IERI_Monograph_Volume_02_Chapter_01.pdf

Vorholzer, A., & von Aufschnaiter, C. (2019). Guidance in inquiry-based instruction-an attempt to disentangle a manifold construct. *International Journal of Science Education, 41*(11), 1562-1577. https://doi.org/10.1080/09500693.2019.1616124

©2026 Copyright by the Science Education Association (Thailand). This article is licensed under CC-BY-4.0.