Critical Artificial Intelligence Literacy and Educational Inequality: Dose–Response Effects in PISA 2025 Peru

[1.] Yonatan Calizaya Ramos (Corresponding Author), [2.] Vilma Valeriana Tapia Ccallo,[3.] Francisco Marino Tipula Mamani,[4]. Ledú Anali Ferreyros Calisaya,[5]. Percy Fredy Quilla Coila,[6]. José Pacori Pacori

[1]Universidad Nacional de Juliaca, Puno, Perú
Email: jcalizayar.doc@unaj.edu.pe
ORCID: https://orcid.org/73-88-73-88
Autor de correspondencia

[2]Universidad Nacional de Juliaca, Puno, Perú
Email: vtapia@unaj.edu.pe
ORCID: https://orcid.org/73-88-73-88

[3]Universidad Nacional del Altiplano, Puno, Perú
Email: ftipula@unap.edu.pe
ORCID: https://orcid.org/73-88-73-88

[4]Universidad Nacional de Juliaca, Puno, Perú
Email: lferreyros@unaj.edu.pe
ORCID: https://orcid.org/73-88-73-88

[5]Universidad Nacional de Juliaca, Puno, Perú
Email: pfquillac.doc@unaj.edu.pe
ORCID: https://orcid.org/73-88-73-88

[6]Universidad Nacional del Altiplano, Puno, Perú
Email: jpacorip.doc@unaj.edu.pe
ORCID: https://orcid.org/73-88-73-88

HOW TO CITE:

Yonatan Calizaya Ramos (Corresponding Author), Vilma Valeriana Tapia Ccallo, Francisco Marino Tipula Mamani, Ledú Anali Ferreyros Calisaya,Percy Fredy Quilla Coila, José Pacori Pacori (2026 Critical Artificial Intelligence Literacy and Educational Inequality: Dose–Response Effects in PISA 2025 Peru International Journal of Special Education, 41(21s), 73 - 88

ABSTRACT:

The expansion of generative artificial intelligence in education raises the need to determine whether school-based opportunities for critical AI literacy improve learning and whether their effects are equitably distributed across students from different socioeconomic backgrounds. This study aimed to estimate the average, dose–response, and heterogeneous effects of school exposure to activities focused on critically evaluating AI-generated information on Peruvian students' science achievement. Microdata from PISA 2025 Peru were used, comprising 6,991 students from 258 schools; the main analysis included 4,966 cases with valid treatment and covariate information. Exposure was measured through the reported frequency of school activities for evaluating the quality of AI-generated information, science achievement through ten plausible values, and socioeconomic, demographic, family, digital-resource, and school-context covariates were incorporated. AIPW-ATT, Propensity Score Matching, Entropy Balancing, and a multivalued Generalized Propensity Score were applied, accounting for PISA weights and socioeconomic heterogeneity. The global ATT was −2.75 PISA points and was not statistically significant. However, frequent exposure yielded a 23.77-point difference in estimated effects between the lowest and highest socioeconomic quartiles (95% CI [5.82, 41.73], p = .009). The findings indicate that critical AI literacy opportunities do not produce generalized average gains, but their effects are distributionally heterogeneous and may be relatively more favorable for socioeconomically disadvantaged students.

Keywords: artificial intelligence, critical literacy, educational inequality, PISA, treatment effects

Introduction

The rapid expansion of generative artificial intelligence (GenAI) is changing the way students search for information, solve problems, and produce knowledge. Recent evidence shows that tools such as ChatGPT can improve academic performance, affective-motivational states, and certain higher-order thinking skills, although effect sizes are heterogeneous and depend on the design of the intervention (Deng et al., 2025). This heterogeneity is consistent with the broader research on AI education, which underscores that technology does not in itself constitute a homogeneous pedagogical intervention (Liu & Zhong, 2024; Yim & Su, 2025). Consequently, the relevant question is no longer only whether AI can support learning, but under what conditions, with what intensity, and for which groups of students its educational opportunities are transformed into observable outcomes.

This conceptual shift requires distinguishing between the use of AI and AI literacy. Long and Magerko (2020) proposed that AI literacy comprises competencies to recognize, understand, use, and critically evaluate intelligent systems. Ng et al. (2021) added dimensions related to use, evaluation, and ethical aspects; Almatrafi et al. (2024) synthesized the literature into six components: recognizing, knowing and understanding, using and applying, evaluating, creating, and ethically navigating. Chiu et al. (2024) also differentiate between knowledge about AI and the competence to apply such knowledge beneficially. Therefore, frequent exposure to generative tools does not necessarily imply that the student possesses competencies to assess the quality, veracity, or relevance of the content produced.

Critical literacy in AI deepens this perspective by placing epistemological evaluation and judgment at the center of the relationship with generative systems. Veldhuis et al. (2025) argue that critical literacy involves questioning assumptions, biases, consequences, and sociotechnical relationships of AI, while Sperling et al. (2024) highlight that practical knowledge must be articulated with epistemological and ethical dimensions. Lintner (2024) also shows that a large part of the available measures are self-reported and that limitations of intercultural validity and content persist. In this research, therefore, the exposure variable is interpreted as a school opportunity to develop critical literacy—to assess the quality of information generated by chatbots—and not as a direct measure of fully acquired individual competence.

The problem is especially important in science education. Performance in science requires interpreting evidence, evaluating explanations, distinguishing reliable information from unsubstantiated claims, and applying critical reasoning. The expansion of generative systems can support these processes, but also facilitate plausible responses without sufficient verification. The development of recurrent opportunities to contrast, question, and validate AI-generated information can, therefore, complement scientific literacy processes. The K-12 literature identifies precisely the need to integrate authentic experiences, assessment activities, and teacher mediation into AI teaching (Casal-Otero et al., 2023; Liu & Zhong, 2024; Yim & Su, 2025).

The distribution of these benefits, however, may be conditioned by previous socioeconomic inequalities. Research on the digital divide has shown that inequality is not limited to material access, but comprises differences in skills, uses, and ability to transform technological resources into outcomes. González-Betancor et al. (2021) showed that socioeconomic origin strongly affects digital access in the household, while school integration of

ICT can act as a compensatory mechanism on frequency and quality of use. More recently, Kastorff et al. (2025), using PISA 2022, found a positive relationship between socioeconomic status and digital competence, but also showed that certain school conditions can reduce such disparities. The emergence of GenAI may be generating a new layer of this inequality (Beckman et al., 2025).

From this perspective, the school can reproduce or compensate for inequalities. When critical literacy opportunities depend mainly on the household, students with higher socioeconomic capital can accumulate advantages; when the school offers structured, continuous, and guided experiences, marginal returns may be greater for students with fewer resources. Beckman et al. (2025) document differentiated literacy profiles and productive use of GenAI; González-Betancor et al. (2021) demonstrate the potential compensatory capacity of the school in the face of digital inequalities; and Kastorff et al. (2025) show that adequate access and inquiry practices can moderate the SES-digital competence association. Therefore, a population-average effect may mask relevant distributional effects.

PISA 2025 offers an exceptional opportunity to analyze this problem. The cycle has science as its main domain and collects information on learning with digital tools and AI. In Peru, the OECD reports that approximately 54% of students use AI chatbots at least weekly to help them learn and that 55.3% report in-class opportunities to evaluate AI-generated information. At the same time, strong socioeconomic inequalities persist in achievement and in AI-related learning opportunities (OECD, 2026). These characteristics make Peru a pertinent scenario to study whether critical AI literacy operates as a new source of inequality or as a potentially compensatory opportunity.

Methodologically, much of the educational evidence on AI is based on descriptive associations, conventional regressions, or localized interventions. In observational data, exposure to a school opportunity is not random: the students treated may differ systematically in socioeconomic status, ICT resources, school trajectory, and institutional context. Propensity score methods allow for improved comparability in observed covariates (Rosenbaum & Rubin, 1983; Stuart, 2010; Austin, 2011), while doubly robust estimators combine treatment and outcome models (Bang & Robins, 2005). Entropy Balancing also allows comparison groups to be recalibrated to reproduce moments of the covariate distribution (Hainmueller, 2012).

An additional limitation of binary approaches is that they reduce intense educational experiences to "exposed/not exposed". Imbens (2000) showed that propensity scores can be extended to multivalued treatments, allowing dose-response relationships to be estimated. In the present context, distinguishing between never, sometimes, frequently and very frequently is substantively important: an occasional activity may represent an isolated experience, while frequent exposure may reflect a more systematic pedagogical integration. In this way, the generalized propensity score (GPS) allows us to analyze whether frequency matters and whether the effects differ by socioeconomic position.

Based on these gaps, the study has three objectives: to estimate the adjusted average effect of school opportunities for critical evaluation of AI-generated information on science performance; to identify a dose–response relationship between four levels of exposure frequency; and to determine whether these effects are heterogeneous between socioeconomically disadvantaged and advantaged students. Three hypotheses are proposed: H1, exposure has a non-zero average effect; H2, potential means of achievement differ between frequency levels; and H3, dose–response effects differ between the first and fourth quartile of the socioeconomic index. The main contribution is to integrate critical literacy in AI, educational inequality, and econometric assessment of heterogeneous effects with national PISA 2025 microdata.

Method

Design and data source

A quantitative, observational, explanatory, and treatment effects assessment study was conducted using secondary microdata from PISA 2025. The Peruvian database, derived from the PISA 2025 International Student File (CY09_MS_STU_PUF.sav), contains 6,991 students from 258 schools. PISA uses a complex sampling design and reports performance using plausible values, so the analysis incorporated the final student weight and, in the main estimates, the available replicated weights. The study did not use personal identifiers and worked exclusively with anonymized public data from the OECD (2026).

The main binary analysis was performed with 4,966 students who had valid information on the treatment variable and the required covariates. Before the common support, 2,778 students were classified as treated and 2,188 as controls. The common support retained 4,963 cases. For the multivalued analysis, the four original treatment categories were preserved; after the trimming procedures, 4,965 observations were used. The variations in analytical size respond to missing information and the support criteria of each specification.

Variables

The main exposure was the frequency with which the student reported having carried out activities in class to evaluate the quality of information generated by artificial intelligence chatbots (PISA item ST436Q16DA). For the binary analysis, "never or almost never" constituted the comparison group and the categories "sometimes", "frequently" and "very frequently" made up the exposed group. For the dose-response analysis, the four original categories were retained. This operationalization represents school opportunities for critical AI literacy and not a direct test of individual competence.

The outcome variable was performance in science, represented by the ten plausible values of PISA (PV1SCIE–PV10SCIE). Each estimate was reproduced for the ten plausible values and the results were combined, avoiding using the individual average as if it were an observed score. Covariates included socioeconomic status (ESCS), household and student ICT resources (ICTRES), sex, age, school grade, repetition, parental education and occupation, household possessions, migratory status, language, urban-rural location, and public-private management.

Table 1

Main variables and analytical function

Dimension	Analytical name (PISA code)	Function	Operationalization
Critical AI Literacy	Frequency of critical evaluation of AI-generated information (ST436Q16DA)	Treatment	Binary and four frequency levels
Science Performance	Ten Plausible Values of Science (PV1SCIE–PV10SCIE)	Outcome variable	PISA score; separate estimate by plausible value
Socioeconomic status	Economic, Social and Cultural Status Index (ESCS)	Confustener and moderator	Continuous and Quartile Weighted
Digital Resources	ICT Resources Index (ICTRES)	Covariable	PISA Standardized Index
Individual features	Sex, age, grade and repetition	Covariates	Previous characteristics of the student
Family context	Parental Education, Parental Occupation, and Home Possessions	Covariates	Cultural and socio-economic resources

School context	Location and management derived from the sample stratum	Secondary heterogeneity	Urban-rural and public-private
Sample design	Final weight of the student and 80 pesos replicated	Weighting and inference	Complex PISA design

Note. Prepared by the author based on PISA 2025. The PISA code ST436Q16DA identifies the frequency of school activities for the critical evaluation of information generated by AI; it is interpreted as a learning opportunity and not as a direct test of competence.

Estimation strategy

For the first objective, the mean effect of treatment on treated patients (ATT), defined as $E[Y(1) - Y(0) \mid T = 1]$, was estimated. The propensity score was estimated from the observed covariates and positivity and common support were verified. The primary estimator was Augmented Inverse Probability Weighting for ATT, selected for its property of double robustness (Bang & Robins, 2005). As robustness tests, Propensity Score Matching 1:1 with replacement and Entropy Balancing were implemented (Hainmueller, 2012). Balance was assessed using standardized mean differences (SMDs), using $|SMD| < .10$ as a practical criterion of adequate balance (Austin, 2011; Stuart, 2010).

For the second objective, a generalized propensity score was estimated using multinomial logit, following the extension of propensity scores to multivalued treatments (Imbens, 2000). The potential means $E[Y(1)]$, $E[Y(2)]$, $E[Y(3)]$ and $E[Y(4)]$ were calculated, as well as the contrasts $2-1$, $3-1$ and $4-1$. For the third objective, the same procedure was estimated separately in Q1 and Q4 of ESCS and then the differences in effects between both groups were contrasted. As complementary analyses, the treatment continuum gradient × socioeconomic level, urban-rural and public-private heterogeneity, and diagnoses of sensitivity to unobserved confounding inspired by Cinelli and Hazlett (2020) and Oster (2019) were estimated.

Analyses were performed in Python using reproducible scripts. Pyreadstat was used for SPSS reading, pandas and numpy for data management, scikit-learn for mapping and matching models, scipy for inference, and matplotlib for visualization. Standard errors, 95% confidence intervals, and bilateral tests with $\alpha = .05$ were reported. Due to the cross-sectional and observational nature of PISA, the results are interpreted as treatment effects adjusted under assumptions of consistency, positivity, and conditional ignorability, and not as equivalent for the purposes of a randomized experiment (Rosenbaum & Rubin, 1983).

Results

Balance and average effect

Before adjustment, treatments and controls presented relevant differences in observed characteristics. The largest standardized mean differences corresponded to ICT resources (.339) and socioeconomic status (.338), followed by urban location (.167) and public management ($|SMD| = .124$). After the ATT weighting, all covariates were below $|.10|$; The ICT resource imbalance was reduced to −.026 and the socioeconomic status imbalance to −.022. The multivalued model showed an equivalent pattern: the maximum previous imbalance of socioeconomic status was .550 and that of ICT resources was .545, while after the multivalued adjustment they were reduced to .031 and .036, respectively. These results indicate a substantial improvement in comparability in observed covariates.

Table 2

Balance of covariates selected before and after adjustment

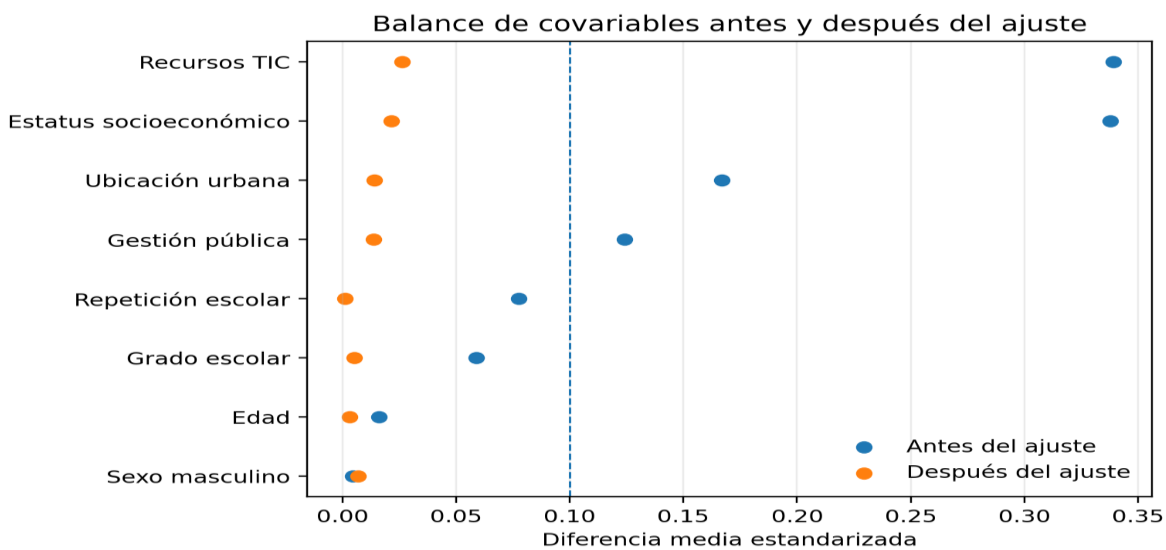
Covariable	SMD before (binary ATT)	SMD After (ATT)	Max. |SMD| Before (GPS)	Max. |SMD| after (GPS)

Socioeconomic status	.338	−.022	.550	.031
ICT Resources	.339	−.026	.545	.036
Urban location	.167	−.014	.250	.045
Public management	−.124	.014	.312	.030
School repetition	−.077	−.001	.187	.039
Male gender	−.004	−.007	.104	.017

Note. Absolute values below .10 after adjustment indicate satisfactory observable balance.

Figure 1

Covariate balance before and after ATT weighting



Note. The reference lines represent the conventional threshold of |SMD| = .10. Prepared by the authors with PISA 2025 microdata. In the original Spanish

The overall ATT estimated by AIPW was −2.75 PISA points (SE = 2.42; 95% CI [−7.49, 1.99]; p = .255). Therefore, no statistically significant evidence of an average gain associated with overall exposure was observed. The direction and magnitude were consistent with PSM (ATT = −1.50) and Entropy Balancing (ATT = −2.72), reducing the possibility that the conclusion depends exclusively on an estimator.

Table 3

Average treatment effect and robustness analysis

Method	Effect (PISA points)	SE	IC95%	p
AIPW-ATT	−2.75	2.42	[−7.49, 1.99]	.255
PSM 1:1 with replacement	−1.50	—	—	—

Entropy Balancing	−2.72	—	—	—

Note. AIPW constitutes the primary estimator; PSM and Entropy Balancing are presented as robustness.

Dose–response relationship

The multivalued GPS showed a nonlinear relationship. The estimated potential mean was 424.71 points for "never/almost never", 418.50 for "sometimes", 426.12 for "frequently" and 428.12 for "very frequently". Regarding the reference level, occasional exposure presented an ATE of −6.21 points (SE = 2.80; 95% CI [−11.70, −0.73]; p = .026). The frequent−never (+1.41; p = .644) and very frequent−never (+3.41; p = .546) contrasts were not statistically significant. The frequency, therefore, does not show a simple monotonic pattern.

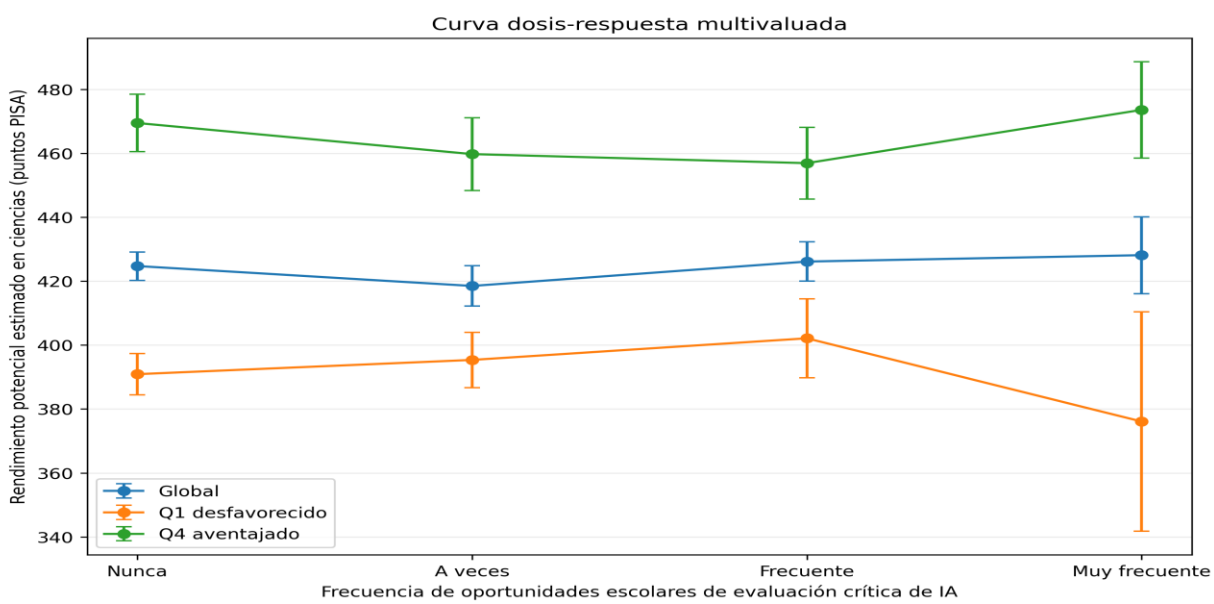
Table 4 *Global Dose–Response Curve of School Critical AI Literacy Opportunities*

Frequency	E[Y(t)]	SE	Contrast vs. Never	p
Never/Almost Never	424.71	2.26	Reference	—
Sometimes	418.50	3.21	−6.21	.026
Frequently	426.12	3.13	+1.41	.644
Very often	428.12	6.13	+3.41	.546

Note. Potential means and contrasts estimated by multivalued GPS/AIPW.

Figure 2

Global dose-response curve and by extreme quartiles of socioeconomic level



Note. The bars represent 95% confidence intervals. Q1 corresponds to the lowest socioeconomic quartile and Q4 to the top. In the original Spanish

Socioeconomic heterogeneity

The main heterogeneity appeared for frequent exposure. In Q1, the frequent−never effect was +11.22 PISA points, while in Q4 it was −12.55. The direct Q1−Q4 contrast was 23.77 points (SE = 9.16; 95%CI [5.82, 41.73]; p = .009). For occasional exposure, the Q1−Q4 difference was +14.18 points (p = .089), while for very frequent exposure it was −18.93 (p = .324). The latter estimate was especially imprecise due to the small number of Q1 observations in the very frequent category (n = 49).

Table 5

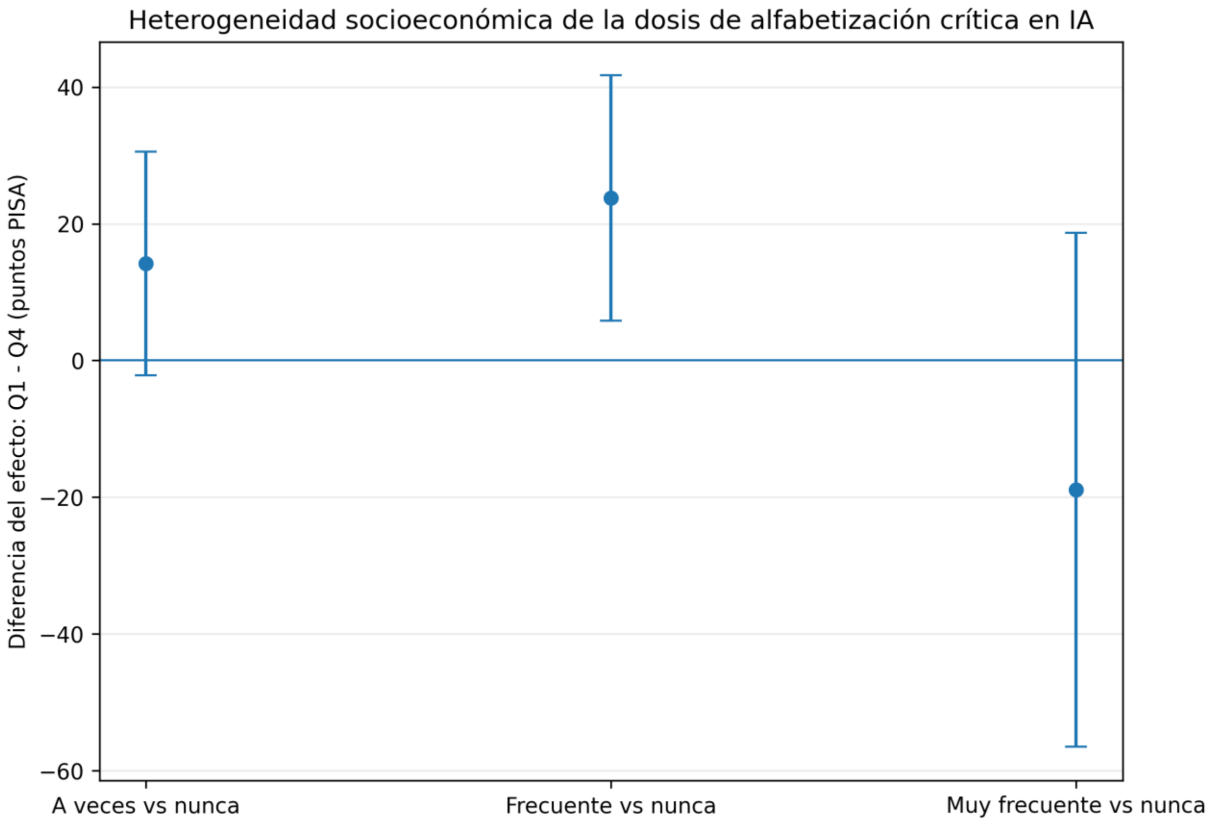
Dose heterogeneity according to socioeconomic status

Contrast	ATE Q1	ATE Q4	Difference Q1−Q4	IC95%	p
Sometimes vs. never	+4.47	−9.72	+14.18	[−2.18, 30.54]	.089
Frequent vs. never	+11.22	−12.55	+23.77	[5.82, 41.73]	.009
Very common vs. never	−14.80	+4.12	−18.93	[−56.52, 18.67]	.324

Note. Q1 = lower socioeconomic quartile; Q4 = upper socioeconomic quartile. The significant contrast is concentrated in frequent exposure.

Figure 3

Q1−Q4 difference in dose-response effects



Note. Positive values indicate a relatively more favorable effect for the lower socioeconomic quartile. Bars: 95% CI. In the original Spanish

Secondary analyses did not show significant heterogeneity by urban-rural context or by public-private management. Urban ATT was −2.22 and rural ATT −4.68; their difference was not significant. Similarly, public ATT was −2.15 and private ATT −4.85. The continuous gradient treatment × socioeconomic status was −3.48 points for each additional standard deviation of ESCS (p = .230), consistent in direction with the Q1–Q4 pattern, but not significant. Sensitivity diagnoses suggest that heterogeneity should be interpreted with caution in view of the possibility of unobserved confounding.

Complementary diagnostics and analyses

Table 6

Descriptive profile of the sample according to school exposure to critical evaluation of information generated by AI

Group	n	Medium socioeconomic level	Science (PISA points)	Urban	Public management
Total	4,966	-1.05	423.16	85.9%	68.4%
Never/Almost Never	2,188	-1.28	418.85	82.6%	71.6%

Sometimes or more	2,778	-0.87	426.64	88.5%	65.8%

Note. Means and percentages incorporate the student's final weight. "Sometimes or more" groups categories together, sometimes, frequently, and very frequently.

Table 7

Average effect on treaties according to socioeconomic quartile

Socioeconomic quartile	THAT	SE	IC 95%	p	Treaties/controls
Q1: Lower socioeconomic status	5.06	4.45	[-3.66, 13.78]	0.256	498/661
Q2	-1.15	3.78	[-8.56, 6.25]	0.760	669/538
Q3	-3.69	4.26	[-12.04, 4.65]	0.386	704/477
Q4: Higher socioeconomic status	-8.21	5.91	[-19.80, 3.37]	0.165	800/436

Note. ATT expressed in PISA points. The quartiles were constructed with the economic, social and cultural status index.

Table 8

Support and effective sample size of the multivalued treatment model

Frequency	n	n Effective GPS	GPS p1	Medium GPS	GPS p99
Never/Almost Never	2,188	1894.5	0.267	0.459	0.691
Sometimes	1,472	1280.4	0.228	0.305	0.360
Frequently	924	765.3	0.078	0.198	0.319
Very often	381	277.9	0.027	0.089	0.176

Note. GPS = generalized propensity score. The effective sample size allows the stability of the weighting at each exposure level to be assessed.

Table 9

Robustness of socioeconomic heterogeneity in the face of alternative definitions of exposure

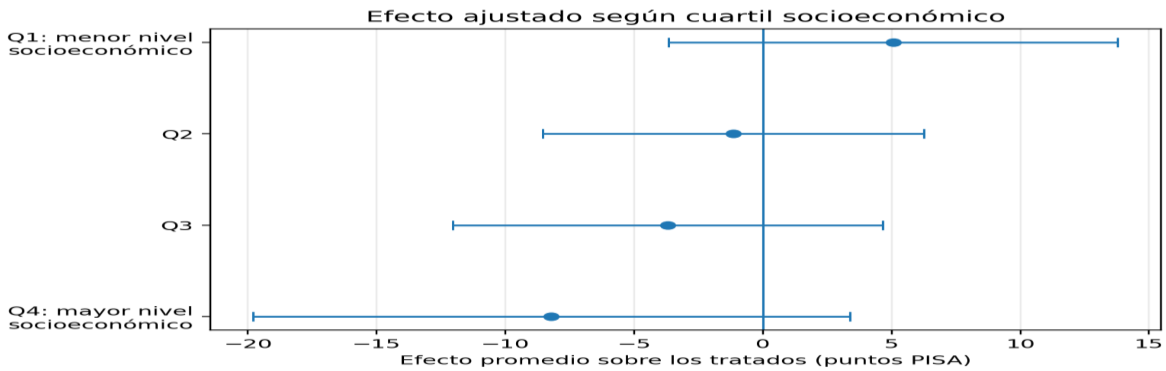
Definition of exposure	THAT global	Difference Q1−Q4	IC 95%	p
Any exposure vs. never	-2.75	13.27	[0.06, 26.49]	0.049
Frequent/very frequent vs. low	4.44	6.44	[-7.73, 20.61]	0.373
Frequent/very frequent vs. never	1.16	11.33	[-5.78, 28.44]	0.194

Very common vs. never	5.48	-8.13	[-39.81, 23.55]	0.615

Note. Q1−Q4 represents the difference between the estimated effect in the bottom and top socioeconomic quartile. Significance is maintained in the primary definition, while alternative specifications show the sensitivity of the result.

Figure 4

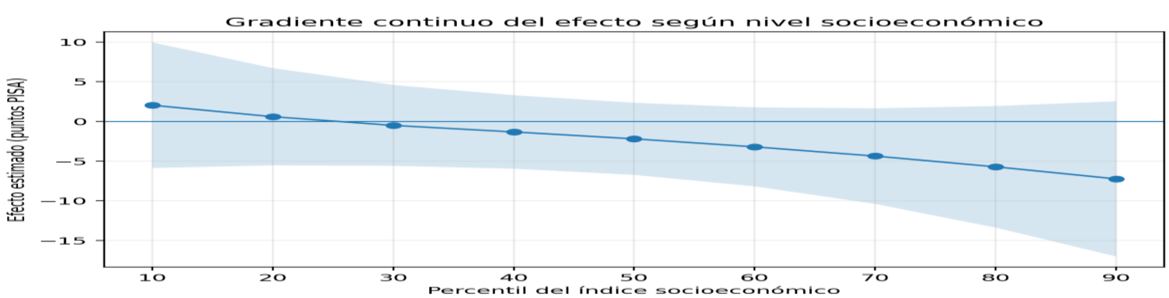
Effect adjusted according to socioeconomic quartile



Note. The dots represent the ATT and the bars represent their 95% confidence intervals. The vertical line indicates null effect. In the original Spanish

Figure 5

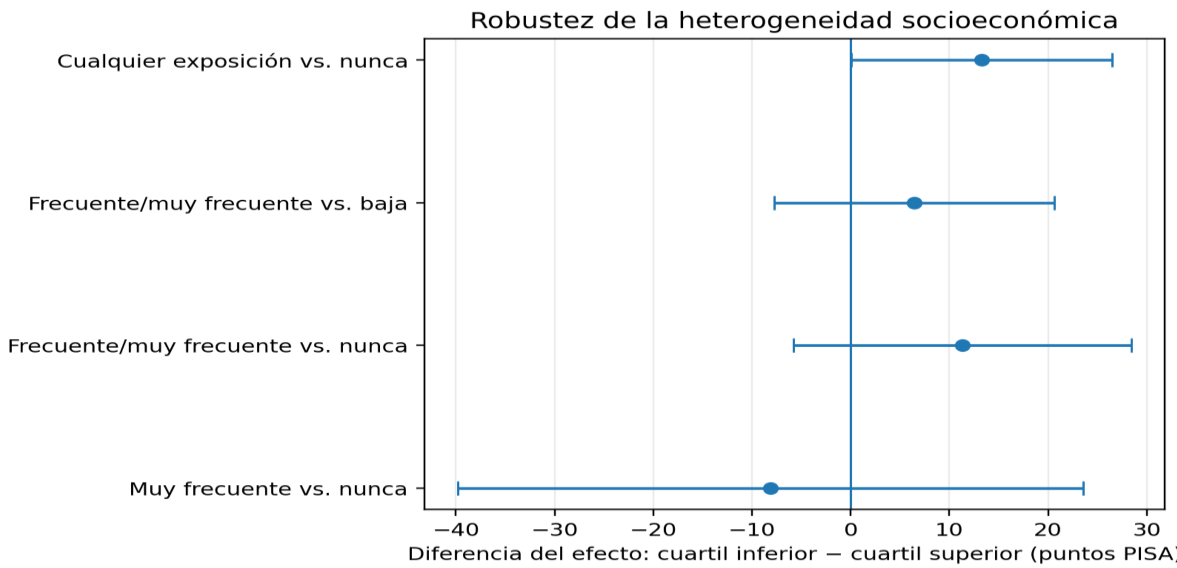
Continuous gradient of the effect according to socioeconomic level



Note. The line represents the estimated percentile effect of the socioeconomic index; the band corresponds to the 95% confidence interval. In the original Spanish

Figure 6

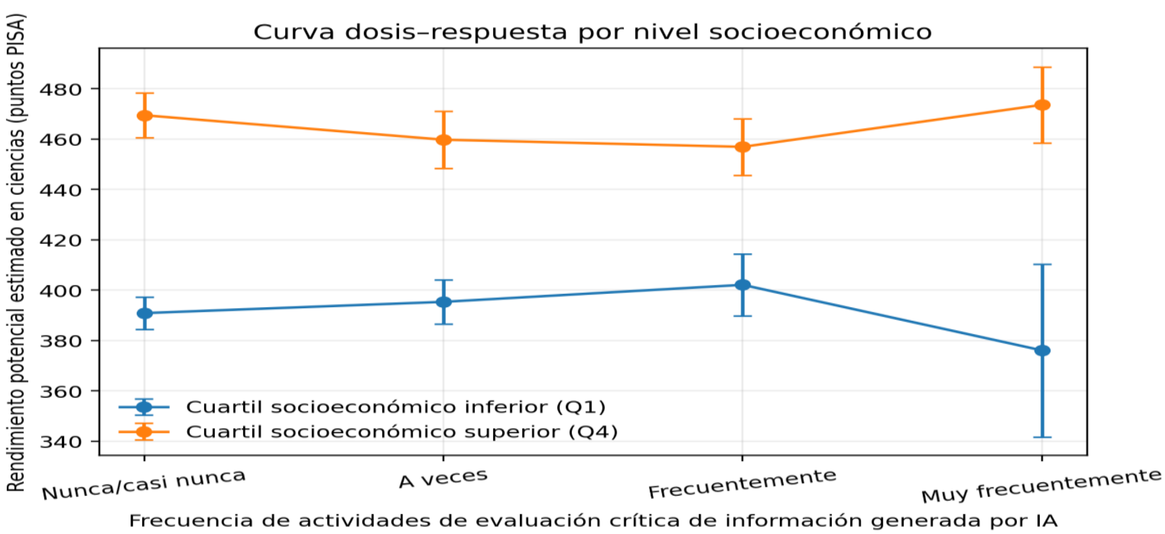
Robustness of socioeconomic heterogeneity in the face of alternative treatment definitions



Note. Positive values indicate a relatively more favorable effect for students in the lower socioeconomic quartile. The vertical line represents no difference. In the original Spanish

Figure 7

Dose–response curve by extreme socioeconomic quartiles



Note. The bars represent 95% confidence intervals. Q1 corresponds to the lowest socioeconomic quartile and Q4 to the top. In the original Spanish

Overall, the complementary diagnoses confirm an adequate balance and show that heterogeneity is concentrated in socioeconomic position rather than in broad institutional categories. The quartile pattern presents a decreasing gradient of effect, although the continuous specification does not reach statistical significance. The alternative definitions of treatment mostly maintain the direction favorable to Q1, but show that the magnitude and significance of the contrast depend on the operationalization of the exposure. Therefore, the finding of 23.77 points for frequent exposure should be interpreted as robust evidence within the multivalued model, but with caution regarding its generalization to other definitions of frequency.

Discussion

The first finding is the absence of a positive average effect of school critical AI literacy opportunities on science performance. Although recent experimental literature tends to identify favorable effects of ChatGPT use on performance and certain higher-order skills (Deng et al., 2025), the present study assesses a distinct exposure: school opportunities to critically evaluate AI-generated information. The conceptual difference is relevant. The positive effects documented in controlled interventions do not imply that every activity associated with AI produces widespread gains. Almatrafi et al. (2024), Chiu et al. (2024), and Lintner (2024) show that AI literacy is multidimensional and that exposure, self-perception, and objective competence are not equivalent.

The fact that the descriptive difference in favor of exposed students disappears after the balance also has a substantive implication. Before adjustment, the groups differed especially in socioeconomic status and ICT resources; after AIPW, PSM, and Entropy Balancing, the effect was close to zero and with a negative direction. This result suggests that an important part of the initially observed advantage was related to observable selection. The literature on digital inequality anticipates this phenomenon: students with higher socioeconomic capital tend to have better resources, competencies, and opportunities to transform technology into educational benefit (González-Betancor et al., 2021; Beckman et al., 2025; Kastorff et al., 2025).

The dose–response analysis also shows that frequency matters, but not in a linear manner. The category "sometimes" presented a lower adjusted potential return than the group that never or almost never performed these activities, while frequent and very frequent exposures did not produce significant average differences. This configuration may be consistent with a pedagogical integration hypothesis: sporadic activities may be insufficient to develop stable routines of verification, contrast, and reasoning, while a higher frequency could reflect more integrated practices. However, the evidence does not support the conclusion that occasional exposure is causally harmful; the pattern should be understood as an adjusted observational relationship that requires replication.

The most relevant result is the socioeconomic heterogeneity of frequent exposure. The contrast of 23.77 PISA points between Q1 and Q4 indicates that the same frequency of school opportunity can have very different estimated returns depending on socioeconomic position. This finding dialogues with the idea of school as a compensatory agent proposed in the literature on digital inequality. González-Betancor et al. (2021) showed that school integration of ICT can moderate inequalities originating in the home; Kastorff et al. (2025) identified that certain school conditions reduce SES gaps in digital competence; and Beckman et al. (2025) warned about a new "GenAI divides" based on differences in literacy and productive use.

A plausible interpretation is that disadvantaged students obtain greater marginal performance from frequent school opportunities for critical evaluation because these opportunities partially replace cognitive, technological, or accompaniment resources that are less available outside of school. This interpretation is consistent with a logic of differentiated marginal performance: for advantaged students, equivalent experiences may provide less novelty because they have greater complementary opportunities. However, the study does not directly measure mediators such as the quality of the teaching scaffolding, domestic access to GenAI, or family practices. Consequently, compensatory interpretation should be formulated as a plausible mechanism and not as demonstrated causal mediation.

The absence of urban-rural and public-private heterogeneity reinforces the idea that broad institutional categories can hide finer differences. Contemporary digital inequality is expressed through skills, resources, trust, pedagogical support, and quality of use, not only through location or school management. Beckman et al. (2025) document very different internal profiles even among students from similar institutions, while Kastorff et al. (2025) show that the specific conditions of availability and use are more informative than the mere presence of technology. For education policy, this suggests avoiding strategies based exclusively on administrative segmentations.

Methodologically, the study shows the usefulness of complementing average estimators with multivalued models. A binary treatment would have produced as a central conclusion a non-significant global ATT and a barely borderline socioeconomic heterogeneity. The GPS revealed that the distributive contrast is concentrated in a specific dose: frequent exposure. This conclusion is aligned with the logic of Imbens (2000), according to which treatment intensities may contain causally relevant information that is lost when dichotomizing. In educational research with AI, where the frequency and depth of use vary markedly, this distinction should be considered in future designs.

The findings have education policy implications. It does not seem justified to simply promote "more AI" or maximize the frequency of use. The potentially relevant intervention is to offer structured opportunities to assess the quality of algorithmic responses: contrasting sources, detecting errors, recognizing biases, and justifying decisions. The frameworks of Almatrafi et al. (2024), Chiu et al. (2024), Veldhuis et al. (2025), and Sperling et al. (2024) converge that adequate literacy should incorporate critical, ethical, and contextual understanding. The results suggest that these practices should especially prioritize students with lower socioeconomic resources, where relative returns may be higher.

The conclusions should, however, be interpreted with caution. PISA is cross-sectional and treatment is self-reported; therefore, unobserved confounding cannot be ruled out. Although AIPW, PSM, Entropy Balancing, and GPS improve the balance in observed covariates, they are not a substitute for random assignment. Cinelli and Hazlett (2020) and Oster (2019) show that sensitivity to omitted variables is an essential component of observational inference. In addition, the "very common" category in Q1 is small in size, which significantly increases uncertainty. Future studies should combine longitudinal data, objective measures of AI literacy, and experimental or quasi-experimental designs.

Overall, the study shifts the focus from "does AI literacy work?" to "for whom, how often, and under what conditions can it work?" The population average showed no benefit, but frequent dosing had a substantially more favorable estimated effect among students in the lower socioeconomic quartile. This suggests that school critical literacy opportunities may have a relevant distributional dimension and that aggregate averages may mask potentially compensatory mechanisms. The contribution to the economics of education lies precisely in identifying this heterogeneity and in showing that digital transformation should be evaluated not only by its average efficiency, but also by its consequences on equity.

Conclusions

The study found no evidence of a generalized average gain in science associated with school critical literacy opportunities in artificial intelligence. The overall ATT was −2.75 PISA points and was not statistically significant, with consistent results in PSM and Entropy Balancing. Therefore, the evidence does not support a simple narrative according to which school exposure to critical AI-related activities homogeneously improves learning.

The main contribution emerges when considering frequency and socioeconomic inequality simultaneously. Frequent exposure, versus never or almost never, had an estimated 23.77 PISA points more favorable effect in Q1 than in Q4 of ESCS. This finding supports the existence of distributive heterogeneity and suggests that certain school opportunities could play a relatively compensatory role for disadvantaged students. However, the effect was not uniform between doses and the very frequent exposure category presented high uncertainty.

In terms of policy, the priority should not be to indiscriminately increase the use of AI, but to design systematic critical literacy practices: source evaluation, contrast, identification of errors and biases, epistemic reasoning, and responsible use. In research, it is recommended to move towards longitudinal, experimental, or quasi-experimental designs, measure objective competence in AI, and analyze the pedagogical quality of the presentation. These steps will make it possible to determine whether the observed compensatory pattern is stable, replicable, and causally sustainable.

Limitations and implications for future research

The study has four main limitations. First, the cross-sectional nature of PISA prevents establishing strong temporality and leaves open the possibility of unobserved confounding. Second, the self-reported frequency of school activities to assess AI-generated information (PISA item ST436Q16DA) measures a learning opportunity and not an objective critical literacy competency. Third, sample sizes are uneven in the categories of the multivalued treatment, especially for very frequent exposure in Q1. Fourth, the Peruvian public base does not offer sufficient subnational geographic variation for spatial econometrics. Future work should incorporate objective AI literacy instruments, more detailed teacher and school information, longitudinal designs, and, when possible, territorial identifiers that allow the study of spatial diffusion and regional inequalities.

Statements

Data availability: The PISA 2025 microdata used in this study is publicly accessible through the OECD. Analysis code: The reproducible Python scripts used to generate the results may be made available as supplementary material. Conflicts of interest: The authors declare no conflicts of interest. Funding: [complete]. Authorship contributions (CRediT): [complete before submission].

References

Almatrafi, O., Johri, A., & Lee, H. (2024). A systematic review of AI literacy conceptualization, constructs, and implementation and assessment efforts (2019–2023). Computers and Education Open, 6, 100173. https://doi.org/10.1016/j.caeo.2024.100173

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Bang, H., & Robins, J. M. (2005). Doubly robust estimation in missing data and causal inference models. Biometrics, 61(4), 962–973. https://doi.org/10.1111/j.73-88.2005.00377.x

Beckman, K., Apps, T., Howard, S. K., Rogerson, C., Rogerson, A., & Tondeur, J. (2025). The GenAI divide among university students: A call for action. The Internet and Higher Education, 67, 101036. https://doi.org/10.1016/j.iheduc.2025.101036

Casal-Otero, L., Catala, A., Fernández-Morante, C., Taboada, M., Cebreiro, B., & Barro, S. (2023). AI literacy in K-12: A systematic literature review. International Journal of STEM Education, 10, 29. https://doi.org/10.1186/s473-88-073-88

Chiu, T. K. F., Ahmad, Z., Ismailov, M., & Sanusi, I. T. (2024). What are artificial intelligence literacy and competency? A comprehensive framework to support them. Computers and Education Open, 6, 100171. https://doi.org/10.1016/j.caeo.2024.100171

Cinelli, C., & Hazlett, C. (2020). Making sense of sensitivity: Extending omitted variable bias. Journal of the Royal Statistical Society: Series B, 82(1), 39–67. https://doi.org/10.1111/rssb.12348

Deng, R., Jiang, M., Yu, X., Lu, Y., & Liu, S. (2025). Does ChatGPT enhance student learning? A systematic review and meta-analysis of experimental studies. Computers & Education, 227, 105224. https://doi.org/10.1016/j.compedu.2024.105224

González-Betancor, S. M., López-Puig, A. J., & Cardenal, M. E. (2021). Digital inequality at home. The school as compensatory agent. Computers & Education, 168, 104195. https://doi.org/10.1016/j.compedu.2021.104195

Hainmueller, J. (2012). Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. Political Analysis, 20(1), 25–46. https://doi.org/10.1093/pan/mpr025

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

Kastorff, T., Müller, M., & Greiff, S. (2025). Digital media use and availability can mitigate the digital divide: Meta-analytic insights from PISA 2022. Computers & Education, 238, 105409. https://doi.org/10.1016/j.compedu.2025.105409

Koch, M. J., Carolus, A., Wienrich, C., & Latoschik, M. E. (2024). Meta AI literacy scale: Further validation and development of a short version. Heliyon, 10(21), e39686. https://doi.org/10.1016/j.heliyon.2024.e39686

Lintner, T. (2024). A systematic review of AI literacy scales. npj Science of Learning, 9, 50. https://doi.org/10.1038/s473-88-073-88

Liu, X., & Zhong, B. (2024). A systematic review on how educators teach AI in K-12 education. Educational Research Review, 45, 100642. https://doi.org/10.1016/j.edurev.2024.100642

Long, D., & Magerko, B. (2020). What is AI literacy? Competencies and design considerations. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1–16). Association for Computing Machinery. https://doi.org/10.1145/3313831.3376727

Ng, D. T. K., Leung, J. K. L., Chu, K. W. S., & Qiao, M. S. (2021). AI literacy: Definition, teaching, evaluation and ethical issues. Proceedings of the Association for Information Science and Technology, 58(1), 504–509. https://doi.org/10.1002/pra2.487

OECD. (2026). PISA 2025 results (Volume I): Future-ready students. OECD Publishing. https://doi.org/10.1787/73451bc5-en

OECD. (2026). PISA 2025 results (Volume I): Peru. OECD Publishing.

Oster, E. (2019). Unobservable selection and coefficient stability: Theory and evidence. Journal of Business & Economic Statistics, 37(2), 187–204. https://doi.org/10.1080/07350015.2016.1227711

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41–55. https://doi.org/10.1093/biomet/70.1.41

Sperling, K., Stenberg, C.-J., McGrath, C., Åkerfeldt, A., Heintz, F., & Stenliden, L. (2024). In search of artificial intelligence (AI) literacy in teacher education: A scoping review. Computers and Education Open, 6, 100169. https://doi.org/10.1016/j.caeo.2024.100169

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical Science, 25(1), 1–21. https://doi.org/10.1214/09-STS313

Veldhuis, A., Lo, P. Y., Kenny, S., & Antle, A. N. (2025). Critical artificial intelligence literacy: A scoping review and framework synthesis. International Journal of Child-Computer Interaction, 43, 100708. https://doi.org/10.1016/j.ijcci.2024.100708

Yim, I. H. Y., & Su, J. (2025). Artificial intelligence literacy education in primary schools: A review. International Journal of Technology and Design Education, 35, 2175–2204. https://doi.org/10.1007/s173-88-09979-w

Zhang, C., Wang, L. H., & Rice, R. E. (2025). U.S. college students' acceptability and educational benefits of ChatGPT from a digital divide perspective. Computers and Education: Artificial Intelligence, 8, 100385. https://doi.org/10.1016/j.caeai.2025.100385