

Before Nonresponse: School-Relative Response Timing and Selective Observability in PISA 2025 Context Questionnaires in Latin America

Jorge Eliécer Villarreal-Fernández

jorgevf2005@gmail.com

Alfonso López Pumarejo Educational Institution <https://orcid.org/0000-0002-4615-8829>

Research Article

Keywords: Doubly robust estimation, Item nonresponse, Large-scale assessment, Paradata, Rapid responding

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11260149/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

Abstract

Context-questionnaire nonresponse in large-scale assessments is usually analyzed after responses become missing, although behavioral change may begin earlier. We examined whether late nonresponse in the 2025 Programme for International Student Assessment was preceded by sustained rapid responding, what factors were associated with that transition, and whether selective loss of respondents affected a study-planning scale and a help-seeking scale located late in the questionnaire.

We analyzed 116,582 students assigned the standard questionnaire across 13 Latin American systems. Rapid responding was identified from module-level timing data and treated as a sustained transition only when it persisted across a qualifying sequence of presented modules. Analyses accounted for the complex survey design and compared students' timing with that of peers in the same school. They also evaluated whether the scales functioned similarly among students who still answered and adjusted population estimates using models of both response probability and expected scale scores. We also examined how conclusions changed under alternative assumptions about missing responses.

About three quarters of students without a later study-planning or help-seeking score had already entered sustained rapid responding before reaching the block containing those scales. Five additional minutes of delay relative to school peers were associated with about 41% higher odds of transition, although the association varied substantially across systems. The association between socioeconomic status and transition was concentrated between schools rather than within them. Among students who still answered, measurement differences were small and localized. Adjustment for selective observability lowered the regional study-planning mean by 0.035 index units, about 0.03 standard deviations (95% confidence interval – 0.055 to – 0.014), but produced little change in help seeking.

Late questionnaire nonresponse can therefore have an observable prehistory. Timing information may help distinguish the quality of observed responses from the separate problem of who remains observable long enough to be measured.

Significance Statement

Long questionnaires do not always lose respondents suddenly. Students may first change how they answer, speed through several sections, and only later stop providing usable responses. Using data from an international student assessment in 13 Latin American education systems, this study shows that such behavioral changes often appeared before later questionnaire nonresponse. Students who fell behind the timing of peers in the same school were more likely to enter sustained rapid responding, and this pattern also varied with school socioeconomic composition.

The consequences were not the same for every questionnaire scale. A scale about planning and monitoring one's studying and a scale about seeking help from classmates and teachers were located late in the questionnaire and showed similar measurement quality among students who still answered. Yet adjusting for who remained observable changed the estimated population mean for study planning

while barely changing help seeking. Missing responses can therefore selectively change which students remain represented in particular measures.

Timing information collected during questionnaires can help detect emerging loss of observability before responses disappear. Such information could support better questionnaire design, monitoring, and sensitivity analyses in large-scale surveys.

1. INTRODUCTION

Context questionnaires in international large-scale assessments are central to analyses of educational opportunity, motivation, self-regulation, and inequality. Their usefulness, however, depends not only on how constructs are measured among respondents but also on who remains observable long enough to provide those measurements. In long computer-administered questionnaires, students may begin by responding normally, later accelerate, and eventually omit items or entire scales. If that process is systematic rather than random, late-questionnaire missingness may be preceded by behavioral signals visible in paradata before the first missing value appears.

Two literatures motivate this possibility. Questionnaire nonresponse is not evenly distributed across an instrument. In PISA 2018, Marcq and Braeken (2025) documented increasing item nonresponse toward the questionnaire's end, with substantial cross-system and individual heterogeneity; South American countries showed especially steep trends. Research on careless and insufficient-effort responding likewise uses response patterns and timing to detect low-quality observed responses. Ulitzsch et al. (2024), for example, developed screen-time-based weights for the estimated probability of careless or insufficient-effort responding. Together, these findings show that response quality can deteriorate within a survey and that timing paradata can capture that deterioration.

A further step is to model disengagement as a transition rather than a fixed respondent characteristic. Welz and Alfons (2023) proposed changepoint methods for identifying individual onset of careless responding. Roman et al. (2024) modeled dynamic shifts into careless or insufficient-effort states, and Xu et al. (2025) used a modified hidden Markov model combining responses and response times. Together, this work shows that behavioral change can be located within a questionnaire. Less clear is how such transitions are structured in clustered survey settings, whether they precede nonresponse, and whether the resulting loss of observability has the same consequences across constructs.

This distinction matters because measurement quality and observability are separate problems. We use selective observability to denote the situation in which the probability of continuing to provide usable responses up to a given scale depends on student characteristics, so that those who are eventually measured no longer represent the full eligible population. A scale may function similarly among respondents who still answer while nevertheless representing a selected subset of the originally eligible population. Conversely, late placement alone does not imply that every scale is equally vulnerable to selection. If the probability of remaining observed is related to characteristics associated with one construct more strongly than another, selective observability may alter one scale estimate while leaving

another largely unchanged. Analyses restricted to observed responses, including conventional measurement-invariance checks, cannot by themselves address that selection process.

The clustered administration of large-scale assessments raises a second issue. Questionnaire burden is usually represented by absolute position or elapsed time, yet students complete questionnaires in social and institutional settings. A student may be far behind same-school peers even when absolute elapsed time is not extreme. Relative delay may therefore capture contextual features missed by absolute time, although mechanisms such as situational pressure or administration practices cannot be identified from timing alone. Separating individual relative timing from the school timing pattern also permits a within-/between-school perspective on disengagement and socioeconomic differences.

Like earlier cycles such as PISA 2018, PISA 2025 releases questionnaire timing data alongside student responses. We use data from 13 Latin American systems to study behavioral nonresponse after separating structural and technical absence. Rather than treating rapid responding as equivalent to missingness, we use it as an observable process indicator and define a sustained transition only when rapid behavior persists across a qualifying sequence of presented modules. Hereafter, transition denotes entry into this pattern, and prior transition denotes a transition that occurred before the block containing the analyzed scales. We then connect this transition to later scale nonresponse, model its timing using both absolute and school-relative measures, and evaluate whether selective observability changes measurement and substantive estimates for two late questionnaire scales, one on planning and monitoring one's study (GOALSET) and one on seeking help from classmates and teachers (SELFREG).

The study makes three contributions. First, it asks whether late nonresponse has an observable prehistory by testing whether sustained rapid-response transitions occur before subsequent scale omission, rather than inferring disengagement from missing values themselves. Second, it evaluates whether a student's delay relative to same-school peers predicts transition conditional on module, the school timing component (the mean deviation of schoolmates from normative pace), baseline response pace, reading performance, socioeconomic composition, and technical-latency proxies, while distinguishing within- from between-school socioeconomic differences. Third, it separates measurement comparability among remaining responders from selective observability in the target population and asks whether the consequences of selection differ across constructs occupying similarly late questionnaire positions.

Accordingly, we address three research questions: (1) How does behavioral nonresponse emerge once structural and technical absence are separated? (2) What factors structure the timing of transition into sustained rapid responding, and does school-relative delay add information beyond individual and school characteristics? (3) Does selective observability alter measurement or population inference for late questionnaire constructs, and are those consequences construct specific? By shifting the analytic starting point from the first missing value to the behavioral process that precedes it, the study treats nonresponse not only as an endpoint but also as the possible culmination of an observable response trajectory.

2. METHODS

2.1 Data Source, Survey Design, and Analytic Populations

We analyzed public-use data from the 2025 Programme for International Student Assessment (PISA) for 13 Latin American systems: Argentina, Brazil, Chile, Colombia, Costa Rica, the Dominican Republic, Ecuador, El Salvador, Guatemala, Mexico, Paraguay, Peru, and Uruguay. PISA uses a two-stage stratified probability design in which schools are first-stage sampling units and eligible students within sampled schools are second-stage units (OECD 2026a). School and student participation rates for each system, calculated according to OECD technical standards (OECD 2025), are reported in Supplementary Table S1. Data collection followed the PISA 2025 main-survey testing period; the international requirement was a 70-day testing window between 1 March and 30 September 2025, unless otherwise agreed (OECD 2026a).

We linked the public-use student file to the OECD questionnaire-timing file. The student file contains questionnaire responses, derived indices, cognitive plausible values, final student weights, and replicate weights. The timing file records total time per questionnaire module in integer seconds and contains neither timestamps nor an administration-order variable. All 120,162 students in the Latin American public-use student file matched exactly one timing record. Of these, 116,582 had the standard student questionnaire, 1,733 were assigned the one-hour (UH) form used for students with special educational needs who could not participate in the regular assessment (Book 199), and 1,847 had no administered student-questionnaire content in the public-use records. Questionnaire-process analyses were restricted to the 116,582 students with the standard form.

Population estimates used the final student weight (W_FSTUWT). Sampling variance used the 80 Fay balanced repeated replication (BRR) weights with Fay coefficient 0.5. Analyses involving cognitive performance incorporated all 10 plausible values for the relevant domain, with imputation variance combined using Rubin's rules (Rubin 1987) unless otherwise stated. Unless otherwise specified, 95% confidence intervals (CIs) used the normal approximation (estimate $\pm$ 1.96 standard errors); the meta-analysis used Hartung-Knapp t-based inference as described in Section 2.5. OECD documentation defined the target population, sampling design, weights, participation standards, and questionnaire wording (OECD 2026a, 2026b).

Analytic populations differed because later analyses required different combinations of questionnaire modules, timing information, socioeconomic measures, school identifiers, and observed scale scores. El Salvador lacked a usable school identifier and Costa Rica lacked the index of economic, social and cultural status (ESCS), so school-level models included 11 systems. Eligibility, exclusions, and numbers of students, schools, and person-module opportunities are reported separately for each analysis.

2.2 Missingness Taxonomy and Rapid-Response Indicator

We classified nonobservation before constructing behavioral indicators because survey data may contain planned omissions, item nonresponse, and failures outside the respondent's substantive answering process (Axenfeld et al. 2024). *Structural nonobservation* comprised valid skips and not-administered items (codes 5 and 7) arising from questionnaire filters or system-level nonadoption. Although PISA permits matrix-sampled questionnaire designs, the analyzed files showed no evidence of student-level matrix sampling: nonadministration patterns were constant within systems and invariant across test booklets. *Technical nonobservation* comprised system-missing values (code 6). It was concentrated in Peru, which accounted for about two-thirds of students with system-missing codes in the target block and was the only system showing a terminal pattern across the final modules; technical loss was therefore treated as censoring. *Behavioral nonresponse* referred to presented items without a substantive response. Rapid responding was analyzed separately as a behavioral process that could precede nonresponse.

Because timing was available only at module level, we adapted the normative-threshold approach with a 10% threshold (NT10; Wise and Ma 2012). For student i in module m ,

$$S_{im} = T_{im} / N_{im}$$

where T_{im} is module time and N_{im} is the number of effectively presented items, including unanswered or invalid responses but excluding structural omissions. The rapid threshold was 10% of the median S_{im} pooled unweighted across the 13 systems among module completers. A module was classified as rapid when S_{im} fell strictly below its threshold. To reduce circular contamination, completers whose rapid-response proportion exceeded 5% across the other modules were excluded and thresholds were iteratively recomputed; classifications stabilized after two iterations. The adaptation uses time per presented item, a regional median, and an iteratively purified reference population. Integer-second timing means that thresholds at or below one second effectively identify zero-second records.

For descriptive analyses, RRG_core was the proportion of core modules presented to each student classified as rapid, excluding optional information and communication technology (ICT) modules; values above 10% indicated elevated rapid responding. Sensitivities used country-specific thresholds, a p99-winsorized mean reference, and an absolute 2-second rule. A 5-second rule was rejected because it classified 7.2% of complete responders as rapid. Full coding and validation appear in Supplementary Appendices S2–S3.

2.3 Defining Transition Into Sustained Rapid Responding

Transition analyses began with 52 *common* modules: modules presented to at least 95% of standard-questionnaire students in every system, allowing terminal technical loss in Peru, and not governed by questionnaire filters or branches (Supplementary Appendix S2). We then excluded the target block (ST432–ST434) and the immediately preceding module ST216, yielding a 48-module clock.

Documentary order was reconstructed from the master questionnaire because the timing file contains durations but no ordering variable.

The event was the first window of five effectively presented modules containing at least three rapid modules. Structurally unadministered modules neither occupied window positions nor censored follow-up; the algorithm advanced to the next presented module. Onset was dated at the first module of the qualifying window, and time-varying covariates were measured at that module. Later modules were required to confirm the 3-of-5 pattern but were not used to construct predictors measured at onset.

Students meeting the criterion in the first window were classified as initially rapid; those whose first qualifying window began at the second through fifth presented module were classified as early transitions. Both groups were excluded from the event-history risk set, which began at the sixth presented opportunity. Eligibility additionally required at least 10 presented modules and a calculable baseline pace from the first five. Intermediate structural gaps did not censor students. Follow-up ended only when fewer than five presented opportunities remained, including after terminal technical loss; the fifth-from-last presented module was therefore the last evaluable onset position. These rules yielded 114,680 students, 18,937 events, and 4,057,242 person-module opportunities. A sensitivity analysis requiring strictly consecutive documentary positions produced nearly identical results in a time-only logistic hazard model (elapsed-time OR 1.254 versus 1.256).

To assess whether the rapid-response transition preceded later nonresponse, we additionally classified students in the standard-questionnaire population, excluding target-block observations affected by technical system missingness. The criterion was whether any qualifying 3-of-5 window was formed entirely from effectively presented clock modules located before the target block in the master questionnaire. Students with fewer than five presented clock modules before the target block were excluded from this classification, leaving 115,017 students eligible for the analysis. We then estimated the weighted probability of missing the official GOALSET or SELFREG score conditional on prior transition, and the probability of prior transition conditional on later score availability. Sampling standard errors used the 80 Fay-BRR replicates. As a sensitivity analysis, onset alone was required to occur before the target block while later modules could confirm the qualifying window.

A two-state hidden Markov model (HMM) was used only to validate persistence, treating rapid classifications as imperfect emissions from engaged and disengaged latent states; details are provided in Supplementary Appendix S4.

2.4 Discrete-Time Event-History Models

Let h_{ij} denote the probability that student i transitioned at presented module j , conditional on not having transitioned previously.

2.4.1 Absolute Time, Reading Performance, and Effort

We first fitted a complementary log-log model,

$$\log[-\log(1 - h_{ij})] = \alpha_j + \lambda_{c(i)} + \beta_1 A_{ij} + \beta_2 P_i + \beta_3 Q_i + \beta_4 F_i + \beta_5 M_i$$

with module and system fixed effects. A_{ij} was cumulative observed questionnaire time before module j , in five-minute units, summed over all core modules presented before j in documentary order, including modules outside the 48-module clock. P_i was baseline pace: the median log ratio of observed to normative time per presented item across the first five presented modules, standardized by its interquartile range (IQR); zero-second times were set to 0.5 seconds before logarithmic transformation. Q_i was reading performance in 100-point units, estimated with each of the 10 plausible values. F_i was centered self-reported test effort (EFFORT1) and M_i indicated missing effort. Exponentiated coefficients are reported as hazard ratios (HRs).

2.4.2 School-Relative Timing and Socioeconomic Composition

We decomposed cumulative temporal deviation from normative pace as

$$D_{ij} = A_{ij} - E_{ij} = G_{s(i)j} + \Delta_{ij}$$

where E_{ij} is cumulative expected time across the same previously presented core modules, with module-specific expected time equal to items presented multiplied by the regional normative median time per item. $G_{s(i)j}$ is the school timing component: the mean deviation among students from the same school in the event-history sample to whom module j was presented and who had not yet transitioned by j . The focal student's relative delay is $\Delta_{ij} = D_{ij} - G_{s(i)j}$. Initially rapid and early-transition students never entered the reference population, and opportunities with no eligible peer were excluded.

The reference was leave-one-out while the focal student still belonged to the peer set. On the event row, the student had already transitioned and was therefore not removed a second time. Restricting peers to modules within their own event-history risk periods produced virtually the same estimate in the independent replication with the first plausible value (OR 1.416 versus 1.417 under the primary definition).

The final model used a logistic link,

$$\text{logit}(h_{ij}) = \alpha_j + \lambda_{c(i)} + \theta_1 \Delta_{ij} + \theta_2 G_{s(i)j} + \mathbf{V}^T \mathbf{X}_i + \theta_3 L_i^W + \theta_4 L_{s(i)}^B$$

Both timing components were expressed per five minutes. $\mathbf{X}_i$ included baseline pace, reading performance, sex, age, grade relative to the system modal grade, and ESCS decomposed into within-school and school-mean components. E_{ij} was not entered separately because, conditional on module fixed effects, it varied primarily through questionnaire routing; observed cumulative time therefore entered through $G_{s(i)j}$ and Δ_{ij} .

Technical latency was proxied by each student's mean time on ST001 and ST003; the school reference was the school median of this measure. The model included the student deviation from the school median (L_i^W) and the school median ($L_{s(i)}^B$), each per five seconds. Costa Rica lacked ESCS and EI

Salvador a usable school identifier, leaving 11 systems. The final analysis included 96,209 students in 3,920 schools, 3,424,112 person-module opportunities, and 15,603 events. Exponentiated coefficients are reported as odds ratios (ORs).

The two links are reported according to the models actually fitted. In the specification without EFFORT used for link-function sensitivity, a logistic refit changed the reading estimate negligibly (OR 0.546 versus clog-log HR 0.548); the school-relative specification was not evaluated under both links.

Regional variances used the 80 Fay-BRR replicates. Computationally intensive regional models used a one-step score update around the full-sample fit; selected complete refits differed by at most approximately 0.001. This approximation was not used for country-specific models. Reading analyses combined the 10 plausible values using Rubin's rules.

Goodness-of-fit diagnostics are reported in Supplementary Appendix S5.4. For clock selection, survey-weighted mean log likelihood favored elapsed time over position (-0.025289 vs. -0.025602 with system fixed effects and an 8-df spline). Conventional i.i.d. AIC/BIC, Hosmer–Lemeshow, and pseudo- R^2 statistics were not interpreted for the survey-weighted Fay-BRR fits; convergence, score-gradient, and exact-refit diagnostics are reported there.

2.5 Cross-System Heterogeneity

We refitted the final school-relative logistic specification within each of the 11 eligible systems. Module fixed effects with no transition events within a system were omitted because their coefficients were unidentified; retaining them prevented replicate-model convergence in Chile, Ecuador, Mexico, and Peru. National coefficients and design variances came from complete refits of all 80 Fay-BRR replicates.

System-specific log odds ratios for five additional minutes of relative delay were synthesized using random-effects meta-analysis with restricted maximum likelihood (REML) and Hartung-Knapp inference. Heterogeneity was summarized as

$$I^2 = (Q - df) / Q$$

(Higgins and Thompson 2002). The prediction interval used a t distribution with $k - 2$ degrees of freedom (Higgins et al. 2009).

The primary latency-adjusted national analysis used the first reading plausible value with exact replicate refits. An earlier analysis combining all 10 plausible values without latency adjustment was retained as a specification sensitivity. Leave-one-system-out analyses assessed influence. An independent R implementation reproduced the overall regional conclusion; discrepancies concentrated in Guatemala and Mexico were treated as specification sensitivity rather than averaged with the primary estimates.

2.6 Measurement Comparability Among Remaining Responders

We evaluated GOALSET (“goal setting”; four items on planning, summarizing, self-questioning, and adapting one’s study approach) and SELFREG (“self-regulation”; four items on seeking help from classmates and teachers) among students who remained observable. For each construct, sustained responders (students with no detected transition) were compared with students who had transitioned into sustained rapid responding but nevertheless produced an official scale score. Group membership was defined from the transition clock, which excludes the target block; all onsets were dated before the block, although confirmation of late qualifying windows could use the final questionnaire module. Exact item wording, response options, and the PUF/master-questionnaire code crosswalk are reported in Supplementary Table S1b.

Measurement comparability was evaluated with multigroup graded-response models under configural, fully invariant, and partial-invariance specifications. Because the transitioned group was much smaller, deterioration in predictive fit was evaluated using survey-weighted, group-specific log-loss, rather than only an overall statistic that could mask misfit in the smaller group. Anchors for partial invariance were selected from a preliminary differential item functioning (DIF) screen using the same data, and alternative anchors were examined as sensitivities. Item-level noninvariance was summarized with ordinal dMACS, an effect-size index of item-level noninvariance.

This analysis addresses comparability only among students who remained sufficiently observable to obtain a scale score; it cannot establish comparability for students who omitted the target block entirely. Full parameters and sensitivity analyses appear in Supplementary Appendix S7.

2.7 Selective Observability and Doubly Robust Estimation

For each scale, the eligible population comprised students assigned the standard questionnaire, excluding target-block observations affected by technical system missingness. Response R_i indicated availability of the official PISA weighted-likelihood estimate (WLE); behavioral nonresponders remained in the target population.

Response probabilities were estimated separately by system using weighted logistic models containing pre-block information: reading performance, ESCS, sex, grade, age, EFFORT1, baseline pace, recent rapid-response behavior, prior transition status, accumulated pre-block delay, selected quadratic terms, and missing-value indicators; exact operational definitions of these pre-block features are given in Supplementary Appendix S8. Estimated propensities were truncated to [0.03, 0.995]. Conditional outcome means among respondents were estimated separately by system using ridge regression with the corresponding predictors.

The primary estimator was the augmented inverse probability weighting (AIPW) estimator

$$\hat{\mu}_{\text{AIPW}} = [\sum_i w_i \{\hat{m}(\mathbf{X}_i) + (R_i / \hat{p}(\mathbf{X}_i)) [Y_i - \hat{m}(\mathbf{X}_i)]\}] / \sum_i w_i$$

where w_i is the final survey weight, $\hat{p}(\mathbf{X}_i)$ the response propensity, and $\hat{m}(\mathbf{X}_i)$ the outcome model. Both nuisance models were re-estimated within each of the 80 Fay-BRR replicates using the first reading

plausible value; between-plausible-value variance from the 10 point estimates was added using Rubin's rules.

We estimated scale means, changes from observed-case means, and ESCS gradients by projecting the AIPW pseudo-outcome on within-system standardized ESCS where available. Regional summaries were calculated both population-weighted and equally weighted across systems. Pairwise system contrasts used Holm multiplicity correction. Inverse probability weighting (IPW), outcome-regression, and alternative propensity-truncation results are reported in Supplementary Appendix S8.

As a construct-specific diagnostic among students who remained observable, we additionally fitted survey-weighted linear models of each WLE on prior transition (as defined in Section 2.3), reading performance, ESCS, sex, grade, and system fixed effects among responders with complete covariate data. Estimates combined the 10 reading plausible values; design variance used 80 complete Fay-BRR refits for the first plausible value plus between-plausible-value variance. Details of the systems retained and both implementations are reported in Supplementary Appendix S8.

AIPW addresses selection on observed pre-block information under conditional missing at random (MAR); it is not an assumption-free recovery of a "true" population mean.

2.8 Nonignorable Nonresponse Sensitivity and Reproducibility

Because conditional MAR is unverifiable, we conducted a pattern-mixture sensitivity analysis around the AIPW estimates to examine missing-not-at-random (MNAR) scenarios. For nonrespondents, the outcome-model component was shifted by

$$\delta\sigma_e$$

where σ_e is the respondent residual standard deviation and δ represents an unobserved departure, in residual-standard-deviation units, from the outcome expected given observed covariates. AIPW estimates were recomputed across a symmetric δ grid, and we report the tipping point at which the adjusted estimate equals the observed-case estimate. Neither the sign nor magnitude of δ is identified by the data, and residual analyses among observed students provided no basis for privileging either direction.

The analysis was sequential and exploratory. Specification decisions, sensitivity analyses, and superseded results were retained in the supplement and reproducibility audit trail. Central findings were reproduced in an independent R implementation written from the analytic specification rather than by translating the original code. Independence therefore refers to separate implementations, not to an independent human analyst. Estimates from the two implementations were not averaged; secondary findings that did not persist across implementations were classified as uncertain.

Analyses were conducted in Python 3.13.5 (pandas 2.2.3, NumPy 2.3.5, scikit-learn 1.8.0) and replicated in R 4.3.3 using haven, data.table, and metafor. Both the original analysis pipeline and the independent R implementation were developed with assistance from large language model tools under the author’s direction. OpenAI ChatGPT (GPT-5.6 Sol) and Anthropic Claude (Claude Opus 5) assisted with code generation and refinement, analytic auditing, organization and visualization of outputs, replication code, and manuscript restructuring. All code was reviewed by the author. The pipeline was executed end to end from the original public-use inputs, and all 22 automated verification checks passed; the execution log is included. The author also verified the numerical results, references, and interpretations and retains full responsibility for the work. Code and replication materials were supplied with the journal submission as supplementary reproducibility materials and will be made public after acceptance.

3. RESULTS

3.1 Nonresponse Emerges From Multiple Mechanisms and a Pre-Missing Transition

Of the 120,162 students in the 13 Latin American systems, 116,582 received the standard student questionnaire, 1,733 received the UH/Book 199 form, and 1,847 had no administered student-questionnaire content in the public-use records. Structural, technical, and behavioral non-observation were empirically distinguishable, with technical loss concentrated in a terminal pattern in Peru.

The timing of sustained rapid responding was strongly nonuniform across the questionnaire. Only 0.395% of students met the 3-of-5 criterion in the initial window and 1.080% first met it at positions 2–5. The corresponding proportions were 3.012% at positions 6–15, 6.631% at positions 16–30, and 5.307% at positions 31–44; 83.575% showed no detected transition. Because these intervals contain different numbers of possible onset positions, their totals should not be interpreted as position-specific hazards. Per position, the weighted onset proportion rose from approximately 0.27% at positions 2–5 to 0.30% at positions 6–15 and 0.44% at positions 16–30, and remained high at approximately 0.38% at positions 31–44, where fewer students remained at risk. After exclusion of initially rapid and early-transition students, the primary event-history risk set contained 114,680 students, 18,937 events, and 4,057,242 person-module opportunities (Fig. 1).

Alt text

Bar chart of weighted onset proportion per possible start position with 95% Fay-BRR confidence intervals. The hatched and separated initial-state bar is 0.395% (0.326–0.464). Transition onset averages 0.270% (0.233–0.307) at positions 2–5, 0.301% (0.271–0.331) at positions 6–15, 0.442% (0.419–0.465) at positions 16–30, and 0.379% (0.355–0.403) at positions 31–44. The last two intervals do not overlap, but these are unconditional onset proportions rather than hazards; fewer students remain at risk at later positions. The hidden Markov inset defines E as engaged and D as disengaged and shows a persistent disengaged state with a much higher probability of rapid responding.

Transition strongly preceded later nonresponse. Among students with a qualifying window formed entirely before the target block, 85.0% subsequently lacked a GOALSET score, compared with 6.2% among those without a prior transition; conversely, 73.2% of students without a GOALSET score had already transitioned before reaching the block. The pattern was nearly identical for SELFREG: 82.0% of students with a prior transition subsequently lacked a score compared with 5.9% without a prior transition, and 73.6% of SELFREG nonresponders had previously transitioned. Results were essentially unchanged when transition onset, rather than the complete qualifying window, was required to occur before the block.

The hidden Markov validation was consistent with a persistent latent change rather than isolated fast responses. The estimated probability of moving from the engaged to the disengaged state was approximately 0.006, whereas persistence in the disengaged state was 0.966. The probability of a rapid response was approximately 0.004 in the engaged state and 0.784 in the disengaged state. Thus, module-level rapid classifications remained noisy indicators, but the latent disengaged state was highly persistent once entered. The HMM served only to validate this interpretation; the observed 3-of-5 rule remained the event definition used in subsequent analyses.

3.2 Relative Timing and School Composition Structure

Transition Risk

In the absolute-time model, each additional five minutes was associated with higher transition risk (HR = 1.246, 95% CI 1.205–1.289). Reading was inversely associated after effort adjustment (HR = 0.553 per 100 points, 95% CI 0.528–0.579), with little change from the model without effort (HR = 0.548). Test effort was weakly inverse, whereas missing effort predicted higher risk (Table 1).

Table 1
Discrete-time models of transition into sustained rapid responding

Panel	Link / effect	Predictor	Estimate (95% CI)
A. Absolute-time model	Complementary log-log / HR	Elapsed questionnaire time (+ 5 min)	1.246 (1.205, 1.289)
A. Absolute-time model	Complementary log-log / HR	Reading performance (+ 100 points)	0.553 (0.528, 0.579)
A. Absolute-time model	Complementary log-log / HR	Self-reported test effort (+ 1 point)	0.980 (0.967, 0.993)
A. Absolute-time model	Complementary log-log / HR	Missing test-effort response	1.470 (1.294, 1.670)
B. School-relative model	Logit / OR	Student-relative delay (+ 5 min)	1.407 (1.356, 1.460)
B. School-relative model	Logit / OR	School timing component (+ 5 min)	0.743 (0.654, 0.844)
B. School-relative model	Logit / OR	Baseline pace (+ 1 IQR)	1.139 (1.089, 1.192)
B. School-relative model	Logit / OR	Reading performance (+ 100 points)	0.639 (0.606, 0.673)
B. School-relative model	Logit / OR	ESCS within school (+ 1 ESCS unit)	1.006 (0.977, 1.035)
B. School-relative model	Logit / OR	School-mean ESCS (+ 1 ESCS unit)	0.746 (0.669, 0.833)
B. School-relative model	Logit / OR	Male	1.132 (1.061, 1.209)
B. School-relative model	Logit / OR	Grade relative to system modal grade (+ 1)	0.889 (0.843, 0.938)
B. School-relative model	Logit / OR	Age (+ 1 year)	1.059 (0.939, 1.195)
B. School-relative model	Logit / OR	Individual latency deviation (+ 5 s)	0.993 (0.986, 0.999)
B. School-relative model	Logit / OR	School latency (+ 5 s)	1.090 (1.031, 1.152)

Notes: Panel A reports hazard ratios from the complementary log-log model; Panel B reports odds ratios from the final school-relative logistic model. Both models include module and system fixed effects. Panel A additionally adjusts for baseline pace, which is not displayed in the table, and includes 114,680 students, 18,937 events, and 4,057,242 person-module opportunities. Panel B includes 96,209 students in 3,920 schools, 15,603 events, and 3,424,112 person-module opportunities across 11 systems. Confidence intervals use the design-based normal approximation. ESCS effects are expressed per one unit of the original ESCS metric, not per standard deviation.

In the school-relative model, five minutes of delay relative to same-school peers still at risk was associated with OR = 1.407 (95% CI 1.356–1.460) after adjustment (Table 1). The school timing component was inversely associated with transition (OR = 0.743, 95% CI 0.654–0.844). This does not imply that slower schools are protective: the model conditions on relative timing, and the school-level association varied across systems.

The association between socioeconomic status and transition operated between rather than within schools. Within-school ESCS was essentially null (OR = 1.006, 95% CI 0.977–1.035), whereas school-mean ESCS predicted lower transition odds (OR = 0.746, 95% CI 0.669–0.833). Reading remained inversely associated (OR = 0.639 per 100 points, 95% CI 0.606–0.673); male students had higher odds (OR = 1.132), and students one grade above the system modal grade had lower odds (OR = 0.889). Latency adjustment did not remove the relative-delay association.

The relative-delay association varied substantially across the 11 eligible systems (Fig. 2), from OR = 1.108 in Argentina to OR = 2.864 in Mexico. The pooled REML–Hartung–Knapp estimate was OR = 1.516 (95% CI 1.257–1.828), with $I^2 = 93.7\%$ and a 95% prediction interval of 0.804–2.860. The independent R implementation gave OR = 1.551 (95% CI 1.264–1.903), $I^2 \approx 93.8\%$, and prediction interval 0.775–3.104. Guatemala and Mexico showed the largest implementation sensitivity; estimates were not averaged.

Alt text

Forest plot for 11 Latin American systems. Most system-specific odds ratios are above one, but their magnitude varies substantially. Mexico and Peru have the largest primary estimates. Crosses for the R implementation are generally close to the primary estimates, with larger discrepancies in Guatemala and Mexico. The pooled odds ratio is above one, while the prediction interval crosses one, indicating strong between-system heterogeneity.

3.3 Measurement Remains Broadly Comparable Between Sustained and Transitioned Responders

Measurement differences between sustained responders and transitioned responders were detectable but small among students who remained sufficiently observable to receive an official scale score (Table 2). For SELFREG, the comparison included 85,820 sustained responders and 3,717 transitioned responders. Imposing scalar invariance increased survey-weighted group-specific predictive log-loss by only 0.001% among sustained responders and by 0.574% among transitioned responders. Partial invariance reduced the transitioned-group loss to 0.137%, recovering 76.2% of the deterioration. Localized dMACS values were 0.104 for Q01 and 0.089 for Q06.

GOALSET showed the same general pattern. The comparison included 83,826 sustained responders and 2,971 transitioned responders. Scalar loss was 0.001% among sustained responders and 0.575% among transitioned responders; after partial invariance, the transitioned-group loss fell to 0.081%, corresponding to 85.8% recovery. The localized dMACS values were smaller than for SELFREG, at 0.034

for Q02 and 0.044 for Q06. These results indicate detectable but localized noninvariance rather than a broad change in what either scale measured among observed responders. They cannot establish comparability for students who omitted the target block entirely.

Table 2
Measurement comparability among sustained and transitioned responders

Scale	n (sustained / transitioned)	Scalar loss (%) (sustained / transitioned)	Partial loss, transitioned (%)	Recovery (%)	Localized dMACS
SELFREG (help seeking)	85,820 / 3,717	0.001 / 0.574	0.137	76.2	Q01 = 0.104; Q06 = 0.089
GOALSET (study planning)	83,826 / 2,971	0.001 / 0.575	0.081	85.8	Q02 = 0.034; Q06 = 0.044
Notes: Loss is the percentage increase in survey-weighted group-specific predictive log loss relative to the configural model. Recovery is the percentage of scalar-model deterioration removed by the partial-invariance specification. dMACS values summarize localized ordinal item noninvariance. The comparison is restricted to students who remained observable enough to receive an official scale score.					

3.4 Selective Observability Has Scale-Specific Consequences

Adjustment for differential observability produced markedly different results for the two late questionnaire scales (Table 3 and Fig. 3). For GOALSET, the population-weighted observed-case mean was 0.0885 and the AIPW mean was 0.0540, yielding a change of -0.0345 (95% CI -0.0552 to -0.0138). The population-weighted average of system-specific standardized shifts was -0.034 standard deviations (SD). The independent R implementation produced a slightly larger but compatible shift of -0.040 (95% CI -0.061 to -0.019). The equally weighted regional estimate was also similar, -0.0362 (95% CI -0.0519 to -0.0205), indicating that the direction was not driven solely by the largest national samples.

Table 3
Consequences of selective observability

Scale (content)	Observed mean	AIPW mean	AIPW shift (95% CI)	Standardized shift (SD)	Prior transition (residual SD; index coefficient, 95% CI)
GOALSET (study planning)	0.089	0.054	-0.035 (-0.055, -0.014)	-0.034	-0.155; -0.152 [-0.22, -0.09]
SELFREG (help seeking)	0.132	0.140	+ 0.008 (-0.011, +0.028)	+ 0.008	+ 0.068; +0.070 [+ 0.01, + 0.13]

Notes: GOALSET is the official index labeled “goal setting” and is described in the manuscript as study planning and monitoring; SELFREG is the official “self-regulation” index and is described as help seeking. AIPW is the primary doubly robust adjustment. In the final column, the first value is the prior-transition difference in residual SD units and the second is its coefficient in index units with 95% confidence interval; both come from the implementation with full design-based inference. The other implementation yielded - 0.152 and + 0.068 residual SD for GOALSET and SELFREG, respectively. The SELFREG association is smaller and its interval lies only narrowly above zero. Independent AIPW replication, ESCS-slope shifts, pairwise contrasts, and MNAR tipping points are reported in the text and Supplementary Materials.

Alt text

Dot-and-whisker plot comparing standardized AIPW shifts for study planning and help seeking across 13 systems. Study-planning shifts are predominantly negative, with the largest reductions in Guatemala and Chile, followed by Paraguay and Brazil. Help-seeking shifts cluster near zero and vary in direction. Confidence intervals are generally wider for system-specific estimates than for the regional summaries.

The system-specific pattern was predominantly negative for GOALSET. The largest standardized AIPW reductions occurred in Guatemala (- 0.119 SD), Chile (- 0.074), Paraguay (- 0.051), and Brazil (- 0.044). This contrasted with SELFREG, for which the observed-case mean was 0.1316 and the AIPW mean was 0.1399. The resulting change, + 0.0083 (95% CI - 0.0110 to 0.0277), was compatible with zero, as was the independent R estimate of + 0.008 (95% CI - 0.010 to 0.025). The population-weighted standardized SELFREG shift was + 0.008 SD, and system-specific changes clustered near zero and varied in direction.

Socioeconomic gradients were more stable than the GOALSET mean. The regional GOALSET ESCS-slope change was - 0.0062 (95% CI - 0.0262 to 0.0137) in the primary analysis and - 0.0019 (95% CI - 0.0217 to 0.0179) in the R replication. For SELFREG, the primary estimate was - 0.0193 (95% CI - 0.0374 to - 0.0012), whereas the replication yielded - 0.0157 (95% CI - 0.0330 to 0.0016). Because the latter interval included zero, the SELFREG slope result was classified as uncertain rather than as a replicated finding. Of 78 pairwise system contrasts, seven showed a nominal change at $p < 0.05$ when the change in the contrast itself was tested, but none survived Holm correction.

Among students who remained observable, the adjusted association with prior transition also differed by construct. After adjustment for reading performance, ESCS, sex, grade, and system, transitioned responders scored approximately 0.15 residual SD lower on GOALSET, whereas the SELFREG difference was small and positive (about 0.07 residual SD). A separate implementation reproduced both estimates closely.

Taken together, these results distinguish selective observability from a general measurement failure. SELFREG and GOALSET occupy similarly late portions of the questionnaire and both showed only small localized measurement noninvariance among remaining responders, yet adjustment for who remained observable changed the regional GOALSET mean while leaving SELFREG essentially unchanged.

3.5 Sensitivity to Missing-Not-at-Random Nonresponse

The AIPW estimates rely on conditional missing at random given observed pre-block information. Pattern-mixture sensitivity analysis therefore displaced the unobserved outcome component among nonresponders by δ residual standard deviations (Fig. 4). For GOALSET, the AIPW-minus-observed difference crossed zero at approximately $\delta = +0.182$. For SELFREG, the corresponding tipping point was approximately $\delta = -0.044$. These values quantify the unobserved residual displacement required for the AIPW-adjusted estimate to equal the observed-case estimate; they are not estimates of the actual direction or magnitude of nonignorability.

Alt text

Two approximately linear sensitivity curves show the adjusted-minus-observed mean as the assumed residual displacement among nonresponders ranges from negative to positive values. The GOALSET curve begins below zero at $\delta = 0$ and crosses zero at about $+0.18$ SD. The SELFREG curve begins slightly above zero and crosses at about -0.04 SD. The observed data do not identify which direction of residual displacement is more plausible.

The observed data provided no basis for privileging either direction of δ . In particular, residual analyses after the full MAR model did not show a monotonic negative gradient with prior response behavior: the residual slope with the pre-block proportion of rapid modules was $+0.019$, and the slope with onset intensity was $+0.025$, with a nonmonotonic categorical pattern. The MNAR analysis was therefore interpreted symmetrically rather than used to assume that nonresponders necessarily had outcomes below those predicted under MAR.

Accordingly, the construct-specific contrast remains conditional rather than absolute: under adjustment for observed pre-block information, GOALSET exhibits a small negative selection-related shift whereas SELFREG remains close to its observed-case mean. Departures from MAR remain unidentified and are represented by sensitivity ranges rather than by a further “corrected” population value.

4. DISCUSSION

This study shows that late questionnaire nonresponse can have an observable behavioral history before responses disappear. In PISA 2025 student questionnaires across 13 Latin American systems, sustained rapid responding commonly preceded later score nonresponse: about three quarters of students without a GOALSET or SELFREG score had already met the rapid-transition criterion before reaching the target block. Transition risk also depended on students' timing relative to peers in the same school and on school socioeconomic composition. Finally, the consequences of selective observability differed across constructs: measurement comparability remained broadly intact among students who still responded, yet adjustment for differential observability shifted GOALSET while leaving SELFREG essentially unchanged. Together, these findings shift attention from missing responses alone to the process by which respondents become progressively less observable.

The findings extend related work. Marcq and Braeken (2025) showed that PISA 2018 item nonresponse increased toward the end of the questionnaire and varied across systems and students. Our results add a behavioral sequence: later omission is frequently preceded by a transition in response timing. Ulitzsch et al. (2024) used screen-time information to weight observed responses by estimated careless or insufficient-effort responding; here the complementary question is who remains observable long enough to contribute a scale score. Changepoint studies (e.g., Welz and Alfons 2023) likewise locate disengagement within, rather than across, entire questionnaires. Our contribution is evidence about what structures that transition in a clustered international survey and what follows for substantive scale estimates.

The strongest descriptive result—that rapid transition preceded later nonresponse for most students who eventually lacked a target score—could appear circular if rapid responding and omission were treated as two measurements of the same event. They are not. Transition is derived from timing paradata, whereas nonresponse is defined from questionnaire response records; more importantly, the qualifying transition is constructed from modules preceding the target block. The result therefore establishes temporal ordering between distinct observed signals. It does not show that rapid responding causes subsequent omission. A more defensible interpretation is that later missingness often has an observable prehistory: behavior recorded before the target scale identifies students at sharply elevated risk of subsequently becoming unobservable. The near-identical GOALSET and SELFREG patterns and their stability under an alternative transition definition argue against treating this association as a peculiarity of one scale.

A second contribution concerns relative rather than absolute timing. Students who accumulated additional delay relative to same-school peers had higher odds of entering sustained rapid responding, conditional on module, the school timing component, baseline pace, reading performance, demographic characteristics, socioeconomic composition, and latency proxies. One possible mechanism is situational pressure when respondents perceive that others are further ahead or have begun finishing. The data cannot observe that mechanism directly: there are no session identifiers or timestamps showing when classmates completed the questionnaire. Relative delay should therefore be interpreted as a contextual marker, not as evidence that seeing peers finish causes disengagement. Its marked cross-system

heterogeneity further suggests that administration conditions or other unmeasured contextual features may modify the association.

The socioeconomic decomposition points to a related distinction. Individual ESCS relative to classmates was essentially unrelated to transition, whereas school-mean ESCS was associated with transition risk. Socioeconomic differences in transition risk therefore appeared mainly between schools rather than between students occupying different socioeconomic positions within the same school. This matters for equity analyses because selective observability may cluster in the same institutional environments used to estimate socioeconomic inequalities. The finding should not, however, be read causally: school composition may proxy institutional, technological, organizational, or administration conditions not directly measured here.

The comparison of SELFREG and GOALSET provides a third contribution: vulnerability to selective observability cannot be inferred simply from scale position. Both constructs occur late and showed only small, localized departures from measurement invariance among students who still produced scores. Yet their adjusted means behaved differently. AIPW produced a small but reproducible downward shift in GOALSET and essentially no regional shift in SELFREG. The respondent-only residual comparison points in the same direction: after adjustment for reading, ESCS, sex, grade, and system, transitioned responders scored about 0.15 residual SD lower on GOALSET, whereas the SELFREG difference was small and positive, at about 0.07 residual SD. Given the item content, this contrast is substantively interpretable: prior transition was associated with lower planning and monitoring of one's study, whereas help seeking did not show the same negative pattern, without implying that disengagement causes either construct. The conclusion remains bounded by the fact that only two target scales were examined.

These findings have practical implications for secondary analysts and questionnaire designers. Analysts using late PISA scales should inspect response coverage and the processes associated with remaining observed rather than treat all missingness as exchangeable. Where selection is strongly predictable, observed-case estimates can be reported alongside adjustment and sensitivity analyses. For designers, timing paradata may help identify emerging disengagement before measurement is lost, motivating prospective tests of module order, session-closing procedures, or adaptive interventions. The present study demonstrates diagnostic potential, not the effectiveness of such designs.

Several limitations qualify these conclusions. Questionnaire timing is recorded only in integer seconds and at module level. Public-use data provide neither session identifiers nor continuous timestamps, preventing direct observation of peer completion or testing-room dynamics. EFFORT1 is self-reported and cannot fully represent effort during the questionnaire. The relative-delay association was highly heterogeneous, with Guatemala and Mexico particularly sensitive to implementation details. Measurement-comparability analyses exclude students who provided no target-scale responses and therefore cannot establish invariance for those who became completely unobservable. AIPW depends on conditional MAR given measured pre-block information, while the direction and magnitude of residual MNAR departure remain unidentified. The residual contrast among observed responders cannot

calibrate that unobserved departure because related response-process variables are already included in the MAR models. All reported relationships are observational. The construct-specific conclusion is based on only two late scales and requires replication across a broader set of constructs. Finally, the analysis covers 13 Latin American systems, so generalization to other PISA systems or large-scale surveys remains an empirical question.

The broader implication is that questionnaire nonresponse should not always be treated as an event that begins when a value becomes missing. In these data, loss of observability was often preceded by detectable behavioral change. Studying that prehistory with paradata can help identify when late-questionnaire missingness is largely inconsequential and when it selectively changes what a survey can measure.

5. CONCLUSION

Questionnaire nonresponse may begin before the first missing value appears. Across 13 Latin American systems, sustained rapid responding frequently preceded later scale omission, and its onset was associated with students' timing relative to school peers and with school socioeconomic composition. The consequences were not uniform across constructs: selective observability shifted GOALSET but left SELFREG largely unchanged. These findings support treating observability as a dynamic survey process distinct from measurement quality among respondents. Timing paradata can therefore help identify when late-questionnaire missingness is emerging and when it may matter for population inference.

Declarations

Data and Code Availability

The analyses use public-use data from the OECD Programme for International Student Assessment 2025, specifically the student-questionnaire and questionnaire-timing files together with the accompanying codebook and questionnaire documentation. The public-use files are available from the OECD PISA 2025 Database (<https://www.oecd.org/en/data/datasets/pisa-2025-database.html>). The study does not redistribute respondent-level OECD microdata. Analysis code, derived nondisclosure tables and figure data, variable definitions, and reproducibility documentation were supplied with the journal submission as supplementary reproducibility materials and will be made public after acceptance.

Acknowledgments and Generative-AI Disclosure

This study design and analysis was not preregistered.

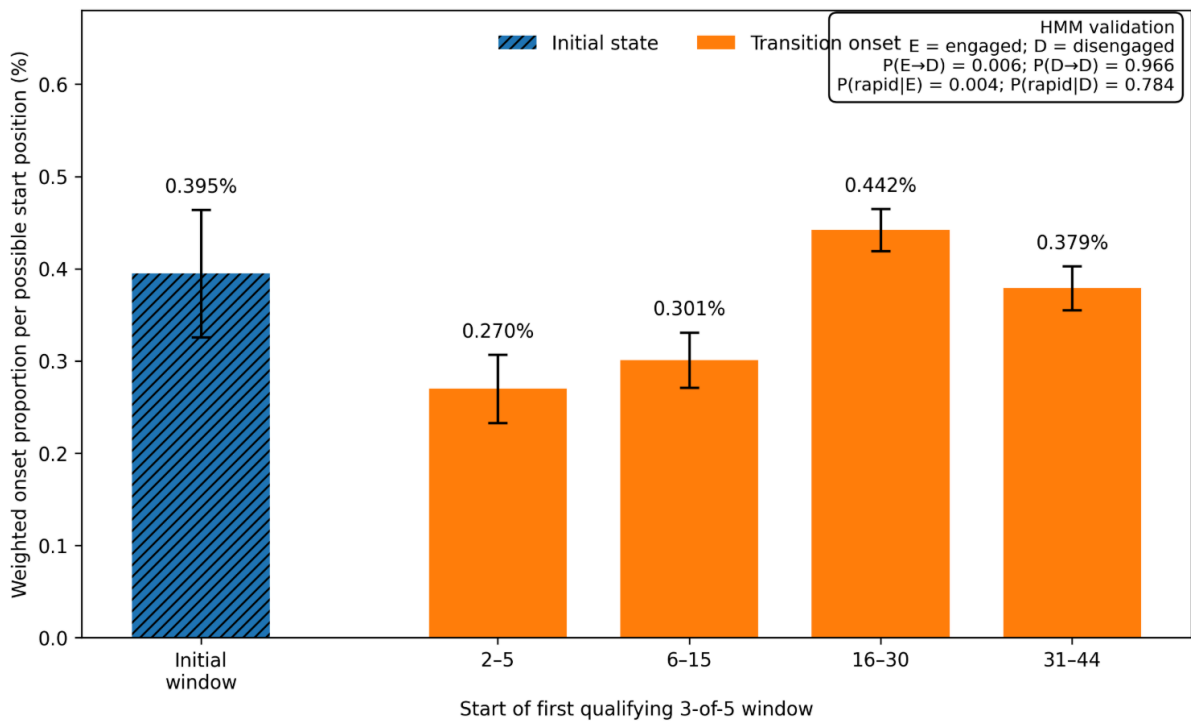
OpenAI ChatGPT (GPT-5.6 Sol) and Anthropic Claude (Claude Opus 5) were used as research-assistance tools for code generation and refinement, analytic auditing, separate R reimplementation from written specifications, organization and visualization of results, and manuscript restructuring and language

revision. The author directed the analytic specifications, reviewed the generated outputs, verified the analyses, and retains full responsibility for the methods, results, interpretations, and final text.

References

1. Axenfeld JB, Bruch C, Wolf C, Blom AG (2024) The Performance of Multiple Imputation in Social Surveys with Missing Data from Planned Missingness and Item Nonresponse. *Surv Res Methods* 18:137–151. 10.18148/srm/2024.v18i2.8158
2. Higgins JPT, Thompson SG (2002) Quantifying Heterogeneity in a Meta-Analysis. *Stat Med* 21:1539–1558. 10.1002/sim.1186
3. Higgins JPT, Thompson SG, Spiegelhalter DJ (2009) A Re-Evaluation of Random-Effects Meta-Analysis. *J Royal Stat Society: Ser (Statistics Society)* 172:137–159. 10.1111/j.1467-985X.2008.00552.x
4. Marcq K, Braeken J (2025) Where Nonresponse Is at Its Loudest: Cross-Country and Individual Differences in Item Nonresponse Across the PISA 2018 Student Questionnaire. *Int J Test* 25:1–24. 10.1080/15305058.2024.2377550
5. OECD (2025) PISA 2025 Technical Standards, Paris: Author. Available at https://www.oecd.org/content/dam/oecd/en/about/programmes/edu/pisa/publications/technical-standards/PISA_2025_Technical_Standards.pdf
6. OECD (2026a) PISA 2025 Results (Volume I): Future-Ready Students. OECD Publishing, Paris. 10.1787/73451bc5-en
7. [dataset] OECD (2026b) PISA 2025 Database, OECD. <https://www.oecd.org/en/data/datasets/pisa-2025-database.html>
8. Roman ZJ, Schmidt P, Miller JM, Brandt H (2024) Identifying Dynamic Shifts to Careless and Insufficient Effort Behavior in Questionnaire Responses; A Novel Approach and Experimental Validation. *Struct Equation Modeling: Multidisciplinary J* 31:775–793. 10.1080/10705511.2024.2304816
9. Rubin DB (1987) *Multiple Imputation for Nonresponse in Surveys*. Wiley, New York. 10.1002/9780470316696
10. Ulitzsch E, Shin HJ, Lüdtke O (2024) Accounting for Careless and Insufficient Effort Responding in Large-Scale Survey Data—Development, Evaluation, and Application of a Screen-Time-Based Weighting Procedure. *Behav Res Methods* 56:804–825. 10.3758/s13428-022-02053-6
11. Welz M, Alfons A (2023) When Respondents Don't Care Anymore: Identifying the Onset of Careless Responding. *arXiv* 230307167. 10.48550/arXiv.2303.07167
12. Wise SL, Ma L (2012) Setting Response Time Thresholds for a CAT Item Pool: The Normative Threshold Method, paper presented at the Annual Meeting of the National Council on Measurement in Education, Vancouver, Canada. Available at <https://www.nwea.org/resources/setting-response-time-thresholds-cat-item-pool-normative-threshold-method/>

Figures



Grouped intervals are divided by the number of possible onset positions; values are unconditional proportions, not hazards. Error bars are 95% Fay-BRR confidence intervals.

Figure 1

Timing of transition into sustained rapid responding. Weighted onset proportion for the first qualifying 3-of-5 rapid-response window, normalized by the number of possible onset positions in each grouped interval. The initial window is shown separately because it represents an initial state rather than an onset transition. Error bars are 95% confidence intervals based on Fay-BRR. Grouped values are unconditional onset proportions per possible start position, not discrete-time hazards. A total of 83.6% of students had no detected transition. The inset reports key parameters from the two-state hidden Markov validation model (E = engaged; D = disengaged).

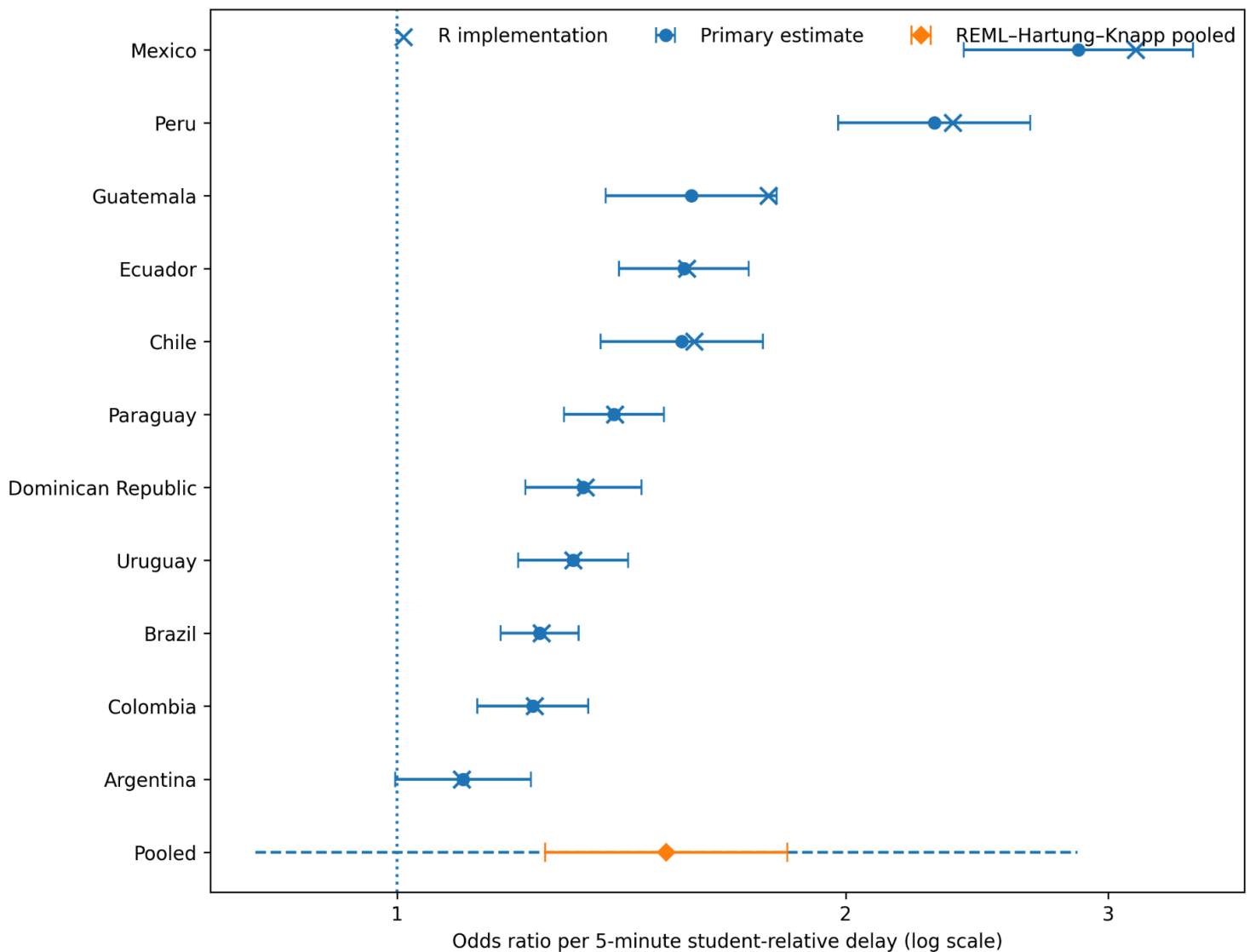


Figure 2

Cross-system heterogeneity in the association between student-relative delay and transition. System-specific odds ratios for transition per five additional minutes of student-relative delay. Circles and confidence intervals show primary exact Fay-BRR estimates; crosses show the independent R implementation. The pooled random-effects estimate uses REML and Hartung-Knapp inference; the dashed horizontal segment on the pooled row is the 95% prediction interval. Guatemala and Mexico show the largest implementation sensitivity.

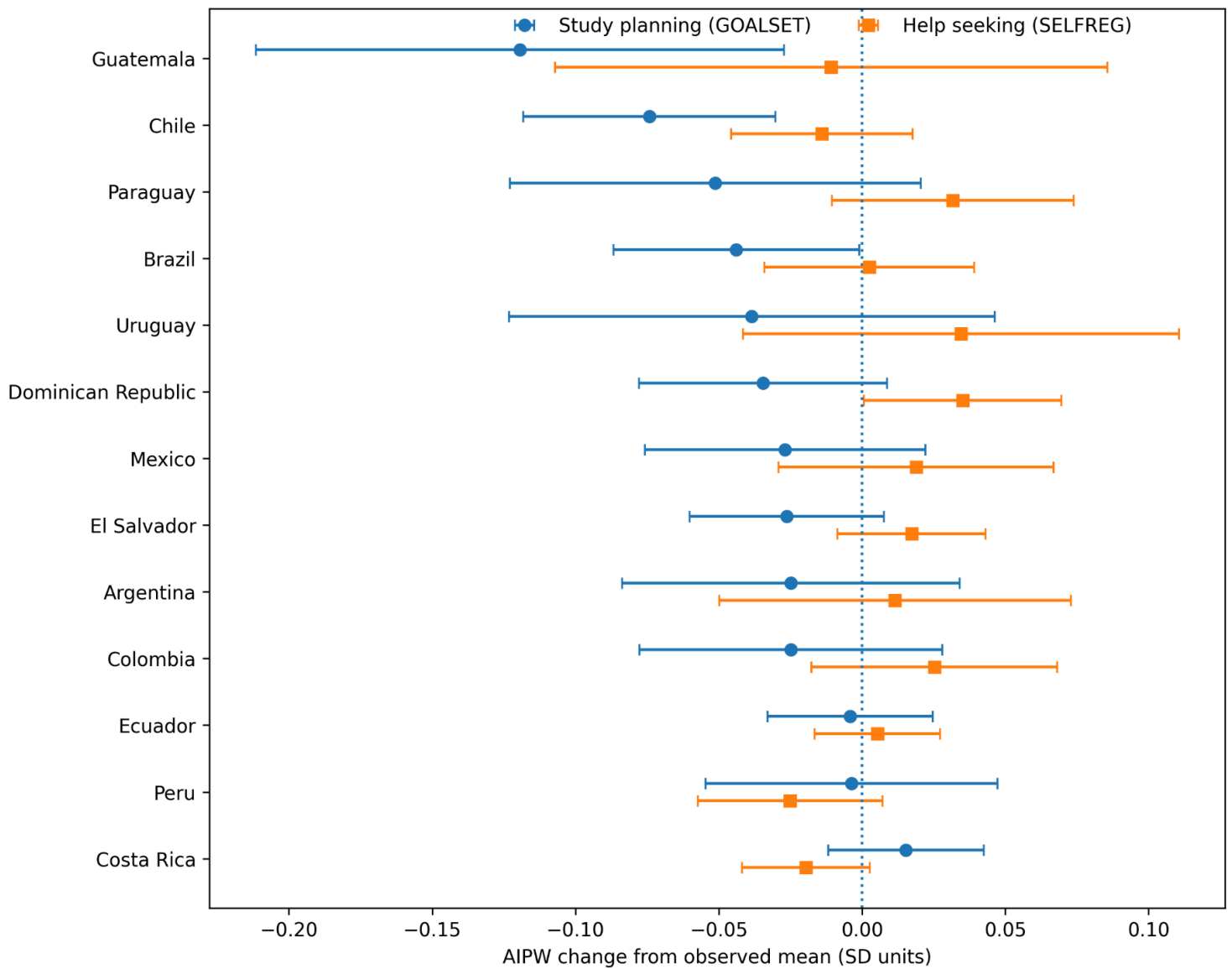


Figure 3

Construct-specific consequences of selective observability. System-specific AIPW changes from observed means, standardized by each system’s observed scale standard deviation. Circles represent study planning (GOALSET) and squares help seeking (SELFREG); horizontal intervals are joint BRR-plus-plausible-value 95% confidence intervals. Negative values indicate that adjustment lowers the estimated mean.

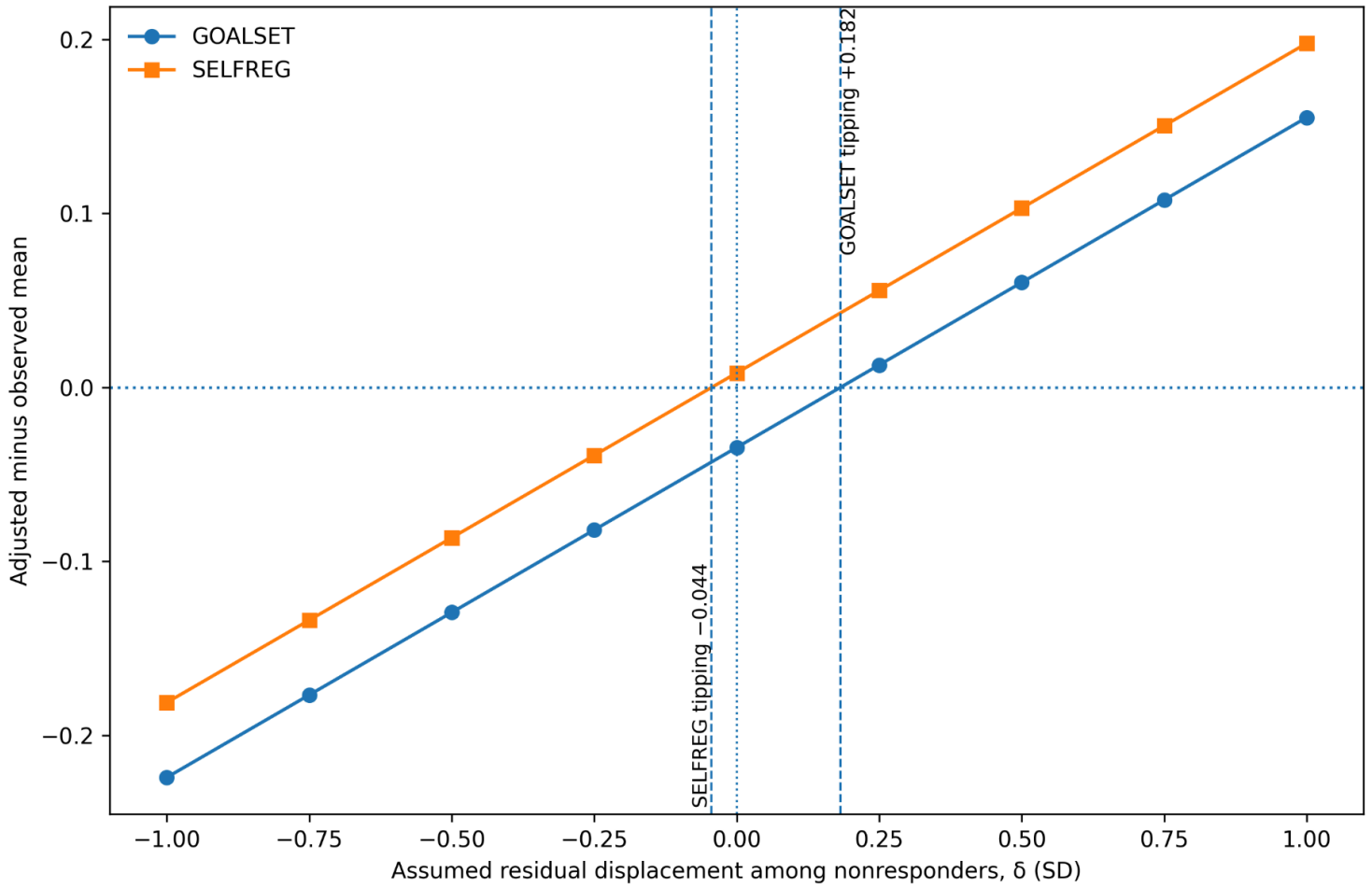


Figure 4

Sensitivity to nonignorable nonresponse. Pattern-mixture sensitivity of population-weighted AIPW shifts to an unobserved residual displacement δ among nonresponders. δ is expressed in respondent residual standard-deviation units. Vertical dashed lines mark the tipping points at approximately $+0.182$ SD for GOALSET and -0.044 SD for SELFREG. The direction of δ is not identified by the observed data.