

How Much of the Engagement–Performance Association Survives Adjustment? A Double Machine Learning Study in Online Higher Education

Eunseon Jo

metacuration@gmail.com

Hanyang University <https://orcid.org/0009-0000-0228-5204>

Hanjin LEE

Handong Global University <https://orcid.org/0000-0002-4068-1336>

Systematic Review

Keywords: behavioral engagement, learning analytics, double machine learning, online higher education, causal inference, academic performance

Posted Date: October 6th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11262205/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

Learning analytics often treats digital activity as an actionable signal of student success, yet how much of a raw engagement–performance association survives adjustment for prior differences remains unclear. Using the Open University Learning Analytics Dataset, we constructed a learner-by-assessment panel of 116,120 continuous assessments from 21,129 students across 22 module presentations. The exposure was pre-submission virtual-learning-environment (VLE) activity intensity, defined in a submission-indexed window that excluded activity after the assessed work was submitted. We estimated a partially linear double machine learning (DML) model with learner-disjoint cross-fitting, examined heterogeneity by prior attainment and assessment format, and conducted module-level, negative-control, overlap, and exposure-definition sensitivity analyses. The unadjusted association was 3.27 assessment points per standard deviation of VLE activity. The DML-adjusted slope was 1.69 points (95% CI [1.58, 1.81]), so 48.2% (95% CI [45.3%, 51.2%]) of the raw association did not survive adjustment. It was 1.68 points larger for tutor-marked open-response assessments than for computer-marked assessments and decreased by 0.17 points per 10-point increase in prior attainment. Deprivation and declared disability showed no detectable moderation. A negative-control outcome remained non-zero, indicating that residual selection cannot be ruled out. Behavioral engagement traces contain useful performance signal, but unadjusted learning-analytics associations can materially overstate an intervention-relevant relationship. For higher-education practice, engagement dashboards and nudging systems should distinguish prediction from causal interpretation and report temporal ordering, adjusted contrasts, and sensitivity to unmeasured confounding.

Introduction

Higher-education institutions increasingly use learning-management-system and virtual-learning-environment traces to identify students who may need support, to prioritize outreach, and to evaluate digital learning provision. Recent work in the *Journal of Computing in Higher Education* illustrates both the promise and the interpretive difficulty of this approach. Aunimo et al. (2025), for example, combined learning-management-system and student-register data to explore factors affecting study duration, whereas Sharma and Lee (2026) showed that high levels of online participation did not necessarily correspond to equally large differences in discourse quality. These studies underscore a central problem for learning analytics: an observable behavioral trace may be highly predictive while remaining ambiguous as evidence about the mechanism that produced the outcome.

This distinction matters once analytics becomes actionable. If a dashboard, advising system, or instructional policy recommends that students should engage more because engagement predicts higher scores, the implied question is no longer merely predictive. It is whether higher engagement itself would be expected to improve performance, after accounting for prior attainment, habitual platform use, enrollment history, and assessment context. Predictive accuracy does not answer that question. As Gašević et al. (2015) argued, learning analytics should remain anchored in learning rather than in the convenience of available data; causal-inference work has subsequently clarified how confounding, post-

treatment adjustment, and temporal ordering can undermine intervention claims drawn from educational logs (Weidlich et al., 2022). Nevertheless, surprisingly little empirical work quantifies how much a familiar engagement–performance association changes when these issues are addressed on a benchmark higher-education corpus.

The Open University Learning Analytics Dataset (OULAD; Kuzilek et al., 2017) provides a useful test bed because it is one of the most widely used open datasets in the field and contains assessed outcomes, learner characteristics, module structure, and fine-grained VLE activity. Reviews of OULAD-based research show that the corpus has been used primarily for prediction, at-risk identification, and related modeling tasks (Alhakbani & Alnassar, 2022; Jin et al., 2024). Those applications are valuable, but they leave an open question: how much of the observed association between VLE activity and assessment performance remains after adjustment for measured pre-exposure history under a design that preserves temporal order?

The Present Study

The present study addresses that question using double machine learning (DML; Chernozhukov et al., 2018). DML is useful here not because machine learning converts observational data into randomized evidence, but because orthogonalization and cross-fitting allow flexible adjustment for a high-dimensional control set while reducing regularization bias in the target coefficient. Crucially, the resulting slope has a causal interpretation only under conditional exchangeability, consistency, positivity, and correct temporal ordering. Because unmeasured confounding cannot be ruled out, we report the result throughout as a DML-adjusted slope and use negative-control and sensitivity analyses to make the remaining uncertainty explicit.

The study is organized around three research questions. RQ1 asks: How much of the unadjusted association between pre-submission VLE activity and assessment performance remains after adjustment for measured pre-exposure history? RQ2 asks: Does the adjusted association vary by prior attainment, assessment format, socioeconomic deprivation, or declared disability? RQ3 asks: How robust is the adjusted association to alternative exposure windows, fold construction, within-corpus module replication, overlap restrictions, and diagnostics for residual confounding? The contribution is therefore not a new prediction algorithm. It is an audit of a commonly actionable learning-analytics relationship, framed around the higher-education decision problem of when a behavioral trace can—and cannot—support intervention-oriented interpretation.

Conceptual and Methodological Framework

Behavioral Engagement as an Instructional Signal

Student engagement is multidimensional and includes behavioral, cognitive, emotional, and agentic components (Mohammadi Zenouzagh et al., 2026). Digital logs capture only a subset of this construct. In particular, click counts and activity records are observable behavioral traces; they do not directly measure cognitive investment, strategy quality, motivation, or knowledge construction. The ICAP framework similarly distinguishes passive, active, constructive, and interactive modes of engagement, which are expected to differ in learning yield even when time on task is similar (Chi & Wylie, 2014). Self-regulated learning theory adds that effective engagement depends on cycles of planning, monitoring, strategy use, and reflection rather than on activity volume alone (Broadbent & Poon, 2015; Zimmerman, 2000).

With this in mind, the present study does not treat VLE activity intensity as the latent construct of engagement. It operationalizes a narrower quantity: the volume of observed VLE activity in a defined window before an assessment is submitted. This distinction is central to construct validity. A positive slope can show that higher behavioral activity is associated with higher assessment performance after adjustment; it cannot, by itself, show that the activity was cognitively productive. The assessment-format analysis provides an indirect test of this issue by asking whether the adjusted association is stronger for tutor-marked open-response work than for computer-marked work, while explicitly recognizing that marking format is not a direct measure of cognitive engagement.

From Prediction to Intervention-Relevant Interpretation

In the potential-outcomes tradition, an intervention-oriented interpretation requires more than prediction: the exposure must be temporally ordered before the outcome, the relevant confounders must be controlled, and sufficient overlap must exist across exposure levels (Imbens & Rubin, 2015; Rubin, 1974). Educational logs complicate each requirement. Students who are already higher attaining may both engage more and score higher; learners with stronger self-regulation may generate more clicks and perform better; and repeated assessments from the same learner create dependence that ordinary row-wise validation can obscure.

DML addresses one part of this problem by estimating a target coefficient after flexibly residualizing both the outcome and the exposure on measured pre-exposure covariates (Chernozhukov et al., 2018). Cross-fitting reduces overfitting bias by estimating nuisance functions on data separate from those used to form the target score. In panels, however, folds should respect the clustering structure; otherwise, another assessment from the same learner can enter the nuisance-training sample (Chiang et al., 2022). We therefore form cross-fitting folds over learners rather than rows and explicitly compare the result with an otherwise identical row-wise split.

Even a well-specified DML estimator cannot remove unmeasured confounding. For this reason, the analysis combines adjustment with falsification and sensitivity. A negative-control outcome asks whether current-window activity predicts a score that was realized before the window opened. A non-zero estimate cannot reflect a forward causal pathway and therefore signals residual selection, reverse

association, or shared causes. In addition, the Cinelli–Hazlett robustness value quantifies how much residual variation an unmeasured confounder would need to explain jointly in the exposure and outcome to reduce the adjusted estimate to zero (Cinelli & Hazlett, 2020). These diagnostics define the boundary of the claim rather than decorating it after the fact.

Method

Design and Data Source

We conducted a secondary analysis of OULAD (Kuzilek et al., 2017), which contains anonymized student records from the Open University in the United Kingdom. The release includes demographic and enrollment information, module and assessment structures, assessment scores, and 10,655,280 VLE interaction records. The analytic unit was one scored continuous assessment submitted by one learner in one module presentation. Sat examinations were excluded because their activity windows and assessment conditions differ from continuous coursework.

The final panel contained 116,120 assessments from 21,129 students across 22 presentations of seven modules. The mean assessment score was 76.11 points ($SD = 18.39$). Table 1 reports the full sequence from the released assessment records to the analytic panel. The study is observational. Its estimand is restricted to submitted, scored continuous assessments that have an earlier assessment in the same presentation and a definable pre-submission exposure window.

Table 1
Construction of the Learner-by-Assessment Analytic Panel

Selection step	Assessments	Students
All assessment records joined to the assessment table	173,912	23,369
Scored (non-missing score)	173,739	23,351
Newly earned (banked scores removed)	171,831	23,291
With a scheduled date	168,966	23,291
Continuous assessment only (sat examinations removed)	166,872	23,285
With at least one earlier assessment in the presentation	141,336	21,245
Definable exposure window	116,860	21,136
Inter-submission gap ≤ 120 days	116,120	21,129
<i>Note.</i> The primary estimand is defined on the final row of the table. Assessments with inter-submission gaps below two days have no definable pre-submission exposure window under the primary rule.		

Table 2
Analytic Panel Characteristics

Quantity	Value
Assessments	116,120
Distinct students	21,129
Module presentations	22
Modules	7
Mean assessment score	76.11
SD of assessment score	18.39
<i>Note.</i> Scores are on a 0–100 scale. <i>SD</i> = standard deviation.	

Exposure, Outcome, and Covariates

Exposure

The primary exposure was $\log(1 + \text{VLE clicks})$ in the interval from the later of 14 days before the learner's actual submission and one day after the previous assessment submission, through the day before the current submission. The exposure was standardized within module presentation. Indexing the window on actual submission rather than the scheduled deadline prevents clicks occurring after the assessed work was submitted from entering the exposure. Opening the window after the previous submission also ensures that the previous-assessment record precedes the exposure in submission time. Assessments submitted fewer than two days after the previous submission have no definable window under this rule and were excluded; inter-submission gaps above 120 days were also excluded as dormant re-entries.

Outcome

The outcome was the institution-recorded score for the continuous assessment on a 0–100 scale. No self-reported learning measure was used.

Assessment Format

OULAD distinguishes tutor-marked assessments (TMAs), which are predominantly open-response tasks, from computer-marked assessments (CMAs), which are predominantly closed-response tasks. We use this distinction descriptively as assessment format. We do not infer that TMA and CMA labels directly measure cognitive engagement. The two formats also differ in score distribution: 28.5% of CMA scores were exactly 100 compared with 0.8% of TMA scores, creating a ceiling that may attenuate associations in the computer-marked regime.

Covariates

The control set was restricted to information available before the exposure window opened. It included demographic and enrollment attributes (gender, region, highest prior education, Index of Multiple Deprivation band, age band, declared disability, credits studied, previous attempts, and registration timing), assessment attributes (assessment type, weight, scheduled date, and position in the assessment sequence), prior attainment (previous score, running mean score, and number of prior assessments), and cumulative VLE activity before the exposure window. Module-presentation indicators captured differences in cohort, grading, and task context. Missing pre-exposure covariates were median-imputed with missingness indicators.

Estimand and Identification Boundary

The primary target was the average partial slope in assessment-score points associated with a one-standard-deviation increase in the observed VLE-activity exposure, conditional on the measured pre-exposure covariates. Under conditional exchangeability, consistency, positivity, and a correctly ordered data-generating process, this parameter can be interpreted as an average partial effect. Those assumptions are not fully testable in observational learning logs. Accordingly, the Results describe the target as an adjusted slope rather than as a proven causal effect.

Figure 1 summarizes the identification logic. Prior attainment, prior VLE activity, learner background, and assessment context may influence both subsequent activity and assessment performance and are therefore controlled using measurements taken before the exposure window. Unmeasured states—such as transient motivation, off-platform study, or task-specific difficulty—may still affect both variables. The negative control and robustness analyses are designed to make that residual uncertainty visible.

Estimation

For RQ1, we estimated a partially linear DML model using gradient-boosted trees for the nuisance functions (Ke et al., 2017). We used five-fold cross-fitting with folds defined over learners, repeated the split five times, and aggregated the target coefficient by the median across repetitions. Standard errors were clustered by learner. For context, we also report the unadjusted ordinary least-squares association and a conventional regression-adjusted coefficient using the same control set. An independent DoubleML implementation (Bach et al., 2022) served as an implementation check. To assess whether within-learner leakage materially changed the result, we reran the same in-repository pipeline using row-wise folds while holding all other settings fixed.

For RQ2, we estimated conditional adjusted slopes with an honest causal forest (Wager & Athey, 2018) and used a linear DML effect-modifier model with learner-clustered covariance for confirmatory contrasts. The prespecified modifiers were prior attainment, assessment format, deprivation rank, and declared disability. Prior attainment was modeled continuously for inference; tercile summaries were used only for descriptive visualization. Between-group conclusions were based on direct interaction contrasts rather than on visual comparison of overlapping confidence intervals.

For RQ3, we applied five robustness checks. First, the full DML analysis was repeated separately within each of the seven modules. Second, the negative-control outcome was the previous assessment score, which was realized before the current exposure window opened. Third, we calculated the Cinelli–Hazlett robustness value for unmeasured confounding. Fourth, we trimmed the first and ninety-ninth percentiles of the exposure residual to examine overlap. Fifth, we re-estimated the model under alternative exposure definitions: a 28-day submission-indexed window, the full inter-submission window, a deadline-indexed truncated window, a legacy untruncated deadline window, and a binary above-median exposure. The alternative definitions are sensitivity analyses, not competing primary estimands.

Analysis Transparency and Reproducibility

The estimand, primary exposure, covariate blocks, effect modifiers, and robustness analyses were recorded in a versioned analysis configuration before the final estimates reported here were produced; the study was not prospectively preregistered. Data-quality auditing identified that deadline-indexed exposure windows could extend beyond learners' actual submission dates. The primary window was therefore re-indexed to actual submissions, and the earlier definitions were retained as sensitivity analyses rather than discarded. This change was based on exposure-timing diagnostics rather than on the exposure–outcome relationship.

All reported estimates can be reproduced from the public OULAD release using Online Resource 1, the accompanying anonymized replication package. The package contains the data-construction scripts, fixed software environment, estimation code, machine-readable outputs used to populate the tables and figures, and a concise specification of the reported analysis. This arrangement is intended to make the inferential choices directly auditable rather than to treat the present coefficient as a universal benchmark.

Use of Generative AI

Generative AI tools were used in three ways. An AI coding assistant (Claude, Anthropic) was used under the authors' direction to write and run parts of the data-construction, estimation, and figure code. Large language model assistants (ChatGPT, OpenAI; Claude, Anthropic) were used for drafting, language editing, and structural revision of the manuscript. An AI assistant was also used to format the manuscript in APA style. The tools did not alter the source data. The authors specified the estimand and the analysis plan, reviewed the code, checked every reported number against the machine-readable outputs in Online Resource 1, checked the citations, and take full responsibility for the submitted work. No AI system is listed as an author.

Results

Sample and Average Adjusted Association

Table 1 shows the construction of the analytic panel, and Table 2 summarizes its main characteristics. The unadjusted association between pre-submission VLE activity and assessment score was 3.27 points

per standard deviation of activity ($SE = 0.077$). Conventional regression adjustment reduced the coefficient to 1.29 points (95% CI [1.18, 1.40]). The DML-adjusted slope with learner-disjoint cross-fitting was 1.69 points ($SE = 0.058$, 95% CI [1.58, 1.81]), equivalent to 0.092 outcome standard deviations (Table 3; Fig. 2).

We summarize the change from the unadjusted to the DML-adjusted coefficient as adjustment attenuation: $(3.27 - 1.69) / 3.27 = 48.2\%$. A learner-clustered delta-method interval placed this attenuation at 45.3% to 51.2%. This quantity is descriptive. It should not be read as the proportion of the original association that was definitively caused by confounding because coefficient attenuation can also reflect nonlinearity, effect weighting under heterogeneity, covariate specification, and measurement error. Notably, the conventional regression-adjusted coefficient was lower than the DML estimate, illustrating that flexible nuisance modeling did not simply remove additional signal.

Changing only the cross-fitting rule from learner-disjoint folds to row-wise folds produced a slope of 1.71 (95% CI [1.59, 1.82]), a 0.7% difference from the primary estimate. The independent DoubleML implementation returned 1.70 (95% CI [1.59, 1.81]). Thus, fold construction was not a major source of bias in this particular specification once rich prior-history covariates were included, although this finding should not be generalized to panels with weaker controls.

Table 3
Association Between Pre-Submission VLE Activity and Assessment Performance

Specification	Estimate	SE	95% CI
Unadjusted association	3.27	0.077	[3.12, 3.42]
Regression-adjusted	1.29	0.057	[1.18, 1.40]
DML-adjusted, learner-disjoint folds	1.69	0.058	[1.58, 1.81]
Same DML pipeline, row-wise folds	1.71	0.057	[1.59, 1.82]
Independent DoubleML implementation	1.70	0.056	[1.59, 1.81]
<i>Note.</i> Estimates are assessment-score points per one standard deviation of $\log(1 + \text{VLE clicks})$ within module presentation. The unadjusted confidence interval is calculated from the reported standard error. DML = double machine learning. VLE = virtual learning environment.			

Heterogeneity by Assessment Format and Learner Characteristics

The adjusted association varied by assessment format and prior attainment (Table 4; Fig. 3). The slope was 1.68 points larger for tutor-marked open-response assessments than for computer-marked assessments ($SE = 0.11$, 95% CI [1.46, 1.90]). It also declined by 0.167 points for each 10-point increase in prior attainment (95% CI [- 0.229, - 0.105]). These two contrasts remained detectable after correction across the four prespecified modifiers.

By contrast, deprivation rank did not show detectable moderation (coefficient = -0.022 per band, 95% CI [$-0.062, 0.018$]), nor did declared disability (difference = 0.234 points, 95% CI [$-0.157, 0.625$]). The absence of detectable moderation should not be interpreted as evidence of equality of effects; the estimates are compatible with small differences that the present design may not resolve.

The forest-based subgroup map was consistent in direction with the confirmatory contrasts. Descriptively, the largest cell was observed for lower-prior-attainment learners on tutor-marked assessments (2.56 points, 95% CI [$1.21, 3.92$]), whereas computer-marked cells were smaller and less precise. However, the forest’s average-effect standard error was substantially wider than the DML average. We therefore interpret the direct contrast tests, not the individual subgroup intervals, as the primary heterogeneity evidence.

Table 4
Prespecified Effect-Modifier Contrasts for the DML-Adjusted Slope

Effect modifier	Difference	SE	95% CI	<i>p</i>
Tutor-marked vs computer-marked	1.680	0.113	[1.459, 1.901]	< .001
Prior attainment, per 10-point increase	-0.167	0.032	[- 0.229, - 0.105]	< .001
Deprivation rank, per band	-0.022	0.020	[- 0.062, 0.018]	.285
Declared disability	0.234	0.200	[- 0.157, 0.625]	.241
Note. The prior-attainment contrast is rescaled to a 10-point increase for interpretability. Positive assessment-format values indicate a larger adjusted slope for tutor-marked than computer-marked assessments. DML = double machine learning.				

Within-Corpus Module Variation

Re-estimating the model separately within each of the seven modules yielded positive adjusted slopes in all modules, ranging from 0.68 to 3.61 points (Fig. 4). The magnitude varied substantially across modules. A descriptive random-effects summary was 1.79 points (Hartung–Knapp 95% CI [$0.82, 2.76$]), whereas the 95% prediction interval extended from -0.59 to 4.17 . Because some learners appeared in more than one module and all modules came from the same provider, these estimates are within-corpus sensitivity analyses rather than independent replications. The wide prediction interval nevertheless underscores that the point estimate should not be transported unchanged to other course contexts.

Robustness and Residual Confounding

The strongest caution came from the negative-control outcome. Current-window activity predicted the previous assessment score by 0.930 points (95% CI [$0.794, 1.065$]) when outcome-derived covariates were excluded from the control set, as required for the falsification check. Because the previous score was realized before the exposure window opened, this non-zero estimate cannot represent a forward effect of current activity. It indicates persistent selection, reverse association, or shared causes that

demographics and prior engagement history do not fully absorb. Adding the prior-attainment block to the primary model reduced the current-score estimate further, but unmeasured time-varying motivation or off-platform study cannot be ruled out.

The Cinelli–Hazlett robustness value was 9.7%, meaning that an unmeasured confounder explaining roughly 9.7% of the residual variance of both the exposure and outcome could reduce the primary adjusted estimate to zero, conditional on the measured controls. This is not a large margin in a behavioral setting. Overlap trimming at the first and ninety-ninth percentiles increased the adjusted slope to 1.96 points, while removing cohort fixed effects left it essentially unchanged at 1.69 points.

The direction of the adjustment was stable across exposure definitions (Fig. 5). The DML-adjusted slope ranged from 1.42 to 2.35 points across the continuous window definitions shown in the figure, and the relative attenuation of the corresponding raw association ranged from 30.6% to 60.3%. The primary submission-indexed 14-day window lay near the middle of this range. Taken together, these analyses indicate that the precise magnitude depends on exposure construction, but the central finding—that the unadjusted association is materially larger than the adjusted relationship—does not depend on a single window definition.

Table 5
Robustness and Sensitivity Analyses

Check	Adjusted slope or value	95% CI	Relative attenuation
Primary 14-day submission-indexed window	1.69	[1.58, 1.81]	48.2%
28-day submission-indexed window	2.14	[2.01, 2.26]	33.7%
Full inter-submission window	1.98	[1.86, 2.11]	30.6%
Deadline-indexed truncated window (superseded)	2.35	[2.22, 2.47]	45.3%
Legacy untruncated deadline window	1.42	[1.31, 1.52]	60.3%
Overlap-trimmed exposure residual (1st–99th pct.)	1.96	[1.83, 2.09]	—
Without cohort fixed effects	1.69	[1.57, 1.80]	—
Negative-control outcome: previous score	0.93	[0.79, 1.06]	—
Cinelli–Hazlett robustness value	9.7%	—	—

Note. Relative attenuation = (unadjusted association – adjusted slope) / unadjusted association for the corresponding exposure definition. The negative-control outcome is the previous assessment score. The robustness value is the share of residual variance in exposure and outcome that an unmeasured confounder would need to explain jointly to reduce the primary estimate to zero. pct. = percentile.

Discussion

What the Adjustment Changes

The first finding is straightforward but consequential: almost half of the raw engagement–performance association did not survive adjustment for measured pre-exposure history. A one-standard-deviation increase in VLE activity corresponded to 3.27 score points before adjustment and 1.69 points after DML adjustment. The surviving relationship remained precisely estimated, but its interpretation is narrower than a conventional learning-analytics dashboard might imply. The data support an adjusted association between a behavioral activity trace and continuous-assessment performance; they do not establish that an institutional intervention that mechanically increases clicks would produce a 1.69-point gain.

This distinction extends prior warnings that learning analytics can drift from learning constructs toward convenient traces (Gašević et al., 2015). It is also consistent with recent JCHE research showing that observable participation patterns can be educationally meaningful without mapping one-to-one onto the quality of learning activity. Sharma and Lee (2026), for example, found highly active online-discussion clusters but only marginal differences in discourse quality. Our results provide a complementary quantitative lesson: activity volume contains signal, but the raw magnitude is highly sensitive to the learner histories that produce both activity and performance.

The negative-control result reinforces this interpretation. Learners with higher current-window activity had scored higher before that window opened, even after adjustment for demographics and earlier engagement history. The primary model reduces this selection by conditioning on lagged attainment, yet the robustness value shows that moderate residual confounding could still erase the adjusted estimate. Thus, the analysis moves the association toward an intervention-relevant quantity but does not complete the identification problem.

Engagement With What, and for Whom?

The heterogeneity results shift the design question from whether engagement matters to where a behavioral-engagement trace is most informative. The adjusted slope was larger for tutor-marked open-response work and smaller for computer-marked work. This pattern is broadly consistent with theories that distinguish more constructive forms of activity from lower-generativity interaction (Chi & Wylie, 2014). Nevertheless, alternative explanations remain plausible. Computer-marked scores had a pronounced ceiling, the surviving CMA sample excludes rapid-fire assessments with no definable exposure window, and the assessment formats may differ in unmeasured content or difficulty. The result should therefore be read as format-specific heterogeneity, not as direct evidence that the trace measured deeper cognitive engagement.

Prior attainment also modified the adjusted relationship: the slope was larger for lower-attaining learners. This finding has practical significance because early-alert and support systems often prioritize students with weaker prior performance. However, the result does not justify indiscriminate engagement

nudges. An intervention may change motivation, strategy, or time allocation in ways that differ from naturally occurring VLE activity. Instead, the finding identifies a subgroup in which experimental tests of targeted support may be especially informative.

No detectable moderation emerged for deprivation or declared disability. Given the uncertainty around subgroup estimates, the appropriate conclusion is not that the association is identical across these groups but that the present dataset did not reveal a stable difference. This is precisely the setting in which direct tests with uncertainty intervals are preferable to visual comparison of subgroup means.

Implications for Higher-Education Learning Analytics

For practice, the main implication is interpretive discipline. Predictive systems can legitimately use VLE activity when the goal is forecasting. Problems arise when a predictive association is translated into an intervention claim without stating the additional assumptions. A dashboard label such as “low engagement risk” may be useful for prioritizing review, but it should not be interpreted as evidence that increasing click volume will necessarily improve learning. Higher-education teams should treat behavioral traces as prompts for diagnosis and conversation rather than as direct measures of effort or learning quality.

For research, four reporting practices would make intervention-oriented claims easier to evaluate. First, define the behavioral exposure precisely and demonstrate that it precedes the outcome. Second, report the raw association alongside the adjusted estimate so that readers can see how much the substantive conclusion depends on measured confounder control. Third, keep repeated observations from the same learner together during cross-fitting, or explicitly show that an alternative fold strategy does not materially change the estimate. Finally, accompany the adjusted estimate with at least one falsification or sensitivity analysis addressing unmeasured confounding. These are reporting practices rather than a proposed formal standard; their purpose is to make the inferential boundary visible and contestable.

The study also demonstrates the value of reproducible public-data audits. Aunimo et al. (2025) showed how institutional LMS and registry data can be combined to study progression in higher education. The present analysis complements that tradition by asking what changes when the same type of behavioral signal is subjected to a stronger identification audit. Because the full pipeline can be rerun from a public source, future work can apply the same logic to other institutions, newer platforms, and experimental interventions without treating the present coefficient as a universal benchmark.

Limitations

Several limitations bound the conclusions. First, the study is observational, and conditional exchangeability cannot be verified. The non-zero negative control and the 9.7% robustness value make residual confounding a substantive rather than merely formal concern. Motivation, off-platform study, contemporaneous life events, and task-specific difficulty are plausible omitted causes.

Second, the exposure is a count-based VLE activity measure. It captures behavioral activity, not the quality, sequence, or cognitive purpose of that activity. The assessment-format contrast is theoretically suggestive but cannot identify cognitive engagement. Third, the panel is conditional on submitted and scored continuous assessments from learners who reached at least a second assessment. Non-submission, early attrition, and final-examination performance are outside the estimand.

Fourth, the window rule improves temporal ordering but removes assessments submitted within a day of the previous submission, disproportionately affecting computer-marked sequences. It also cannot identify when a previous score became visible to the learner or when work on the next submission began. Fifth, the data come from a single distance-learning provider in 2013–2014. Platform design and student behavior have changed, and the module-level variation observed here cautions against transporting the point estimate to contemporary campus, blended, or AI-mediated environments.

Finally, DML reduces regularization bias in the presence of flexible nuisance models; it does not make the exposure randomly assigned. The present analysis should therefore be interpreted as a more defensible observational estimate and an audit of inferential sensitivity, not as a substitute for randomized or quasi-experimental evaluation of specific engagement interventions.

Conclusion

In a large public dataset of online higher education, the familiar association between VLE activity and assessment performance became substantially smaller after adjustment for prior attainment, prior engagement, learner background, and assessment context. The raw slope was 3.27 score points per standard deviation of activity; the DML-adjusted slope was 1.69 points, with 48.2% of the raw association not surviving adjustment. The adjusted relationship was stronger for tutor-marked open-response assessments and for learners with lower prior attainment, but its magnitude varied across modules and remained vulnerable to residual confounding.

The result does not imply that engagement is unimportant. It shows that the meaning of an engagement trace depends on the inferential question. For prediction, VLE activity can be a useful signal. For intervention, a stronger standard is required: temporal ordering, pre-exposure adjustment, learner-aware validation, direct heterogeneity tests, and explicit sensitivity to unmeasured confounding. Taken together, the findings support a cautious but actionable conclusion for higher-education learning analytics: use behavioral traces to identify where further support or investigation may be warranted, but do not equate a predictive association with the effect of increasing engagement.

References

1. Alhakbani HA, Alnassar FM (2022) Open learning analytics: A systematic review of benchmark studies using Open University Learning Analytics Dataset (OULAD). In 2022 7th International

- Conference on Machine Learning Technologies (ICMLT) (pp. 81–86). Association for Computing Machinery. <https://doi.org/10.1145/3529399.3529413>
2. Aunimo L, Kauttonen J, Vahtola M, Huttunen S (2025) Combining data from a LMS and a student register for exploring factors affecting study duration. *J Comput High Educ* 37(3):1032–1055. <https://doi.org/10.1007/s12528-024-09414-4>
 3. Bach P, Chernozhukov V, Kurz MS, Spindler M (2022) DoubleML: An object-oriented implementation of double machine learning in Python. *J Mach Learn Res* 23(53):1–6. <https://www.jmlr.org/papers/v23/21-0862.html>
 4. Broadbent J, Poon WL (2015) Self-regulated learning strategies & academic achievement in online higher education learning environments: A systematic review. *Internet High Educ* 27:1–13. <https://doi.org/10.1016/j.iheduc.2015.04.007>
 5. Chernozhukov V, Chetverikov D, Demirer M, Duflo E, Hansen C, Newey W, Robins J (2018) Double/debiased machine learning for treatment and structural parameters. *Econometrics J* 21(1):C1–C68. <https://doi.org/10.1111/ectj.12097>
 6. Chi MTH, Wylie R (2014) The ICAP framework: Linking cognitive engagement to active learning outcomes. *Educational Psychol* 49(4):219–243. <https://doi.org/10.1080/00461520.2014.965823>
 7. Chiang HD, Kato K, Ma Y, Sasaki Y (2022) Multiway cluster robust double/debiased machine learning. *J Bus Economic Stat* 40(3):1046–1056. <https://doi.org/10.1080/07350015.2021.1895815>
 8. Cinelli C, Hazlett C (2020) Making sense of sensitivity: Extending omitted variable bias. *J Royal Stat Society: Ser B (Statistical Methodology)* 82(1):39–67. <https://doi.org/10.1111/rssb.12348>
 9. Gašević D, Dawson S, Siemens G (2015) Let's not forget: Learning analytics are about learning. *TechTrends* 59(1):64–71. <https://doi.org/10.1007/s11528-014-0822-x>
 10. Imbens GW, Rubin DB (2015) Causal inference for statistics, social, and biomedical sciences: An introduction. Cambridge University Press. <https://doi.org/10.1017/CBO9781139025751>
 11. Jin L, Wang Y, Song H, So H-J (2024) Predictive modelling with the Open University Learning Analytics Dataset (OULAD): A systematic literature review. In A. M. Olney, I.-A. Chounta, Z. Liu, O. C. Santos, & I. I. Bittencourt (Eds.), *Artificial intelligence in education. Posters and late breaking results, workshops and tutorials, industry and innovation tracks, practitioners, doctoral consortium and blue sky* (pp. 477–484). Springer. https://doi.org/10.1007/978-3-031-64315-6_46
 12. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, Ye Q, Liu T-Y (2017) LightGBM: A highly efficient gradient boosting decision tree. *Adv Neural Inf Process Syst* 30:3146–3154
 13. Kuzilek J, Hlosta M, Zdrahal Z (2017) Open University Learning Analytics dataset. *Scientific Data*, 4, Article 170171. <https://doi.org/10.1038/sdata.2017.171>
 14. Mohammadi Zenouzagh Z, Admiraal W, Saab N (2026) Empowering student engagement: The dynamics of learner traits in digital feedback environments. *J Comput High Educ* 38(2):678–703. <https://doi.org/10.1007/s12528-025-09459-z>
 15. Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psychol* 66(5):688–701. <https://doi.org/10.1037/h0037350>

16. Sharma P, Lee M (2026) Rich clubs and engagement in online discussions: Implications for participatory learning design. J Comput High Educ Adv online publication.
<https://doi.org/10.1007/s12528-026-09503-6>
17. Wager S, Athey S (2018) Estimation and inference of heterogeneous treatment effects using random forests. J Am Stat Assoc 113(523):1228–1242. <https://doi.org/10.1080/01621459.2017.1319839>
18. Weidlich J, Gašević D, Drachsler H (2022) Causal inference and bias in learning analytics: A primer on pitfalls using directed acyclic graphs. J Learn Analytics 9(3):183–199.
<https://doi.org/10.18608/jla.2022.7577>
19. Zimmerman BJ (2000) Attaining self-regulation: A social cognitive perspective. In M. Boekaerts, P. R. Pintrich, & M. Zeidner (Eds.), Handbook of self-regulation (pp. 13–39). Academic Press.
<https://doi.org/10.1016/B978-012109890-2/50031-7>

Figures

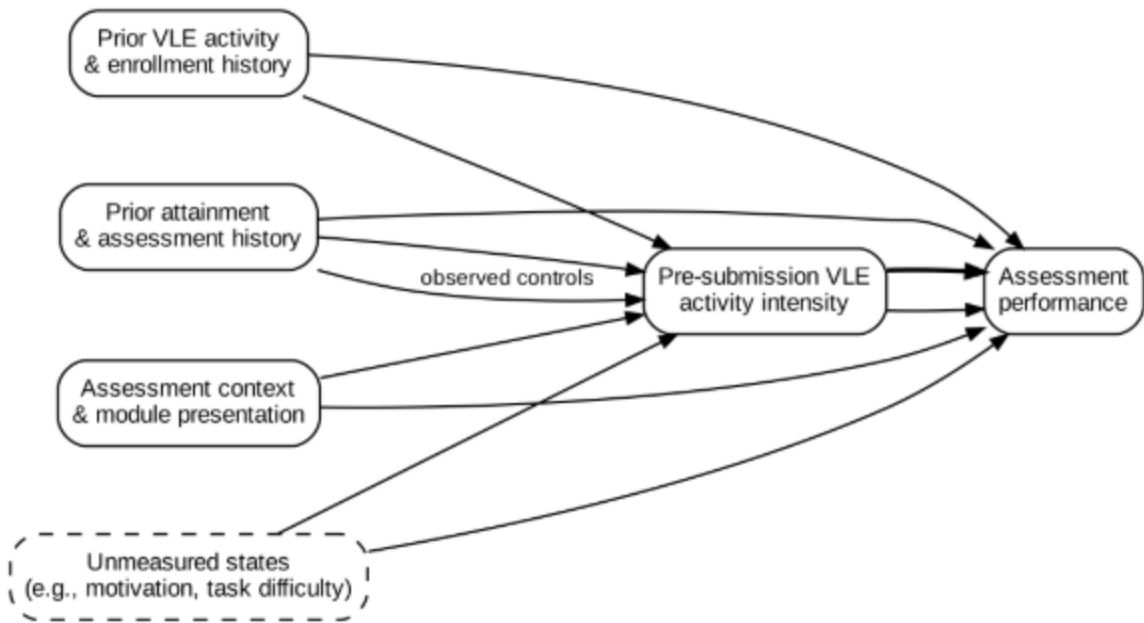


Figure 1

Identification Logic for the Observational Learning-Analytics Analysis

Note. Solid boxes denote observed pre-exposure variables; the dashed box represents plausible unmeasured states that can still confound the adjusted relationship. The emphasized arrow is the relationship of interest, not an assumption of random assignment. VLE = virtual learning environment.

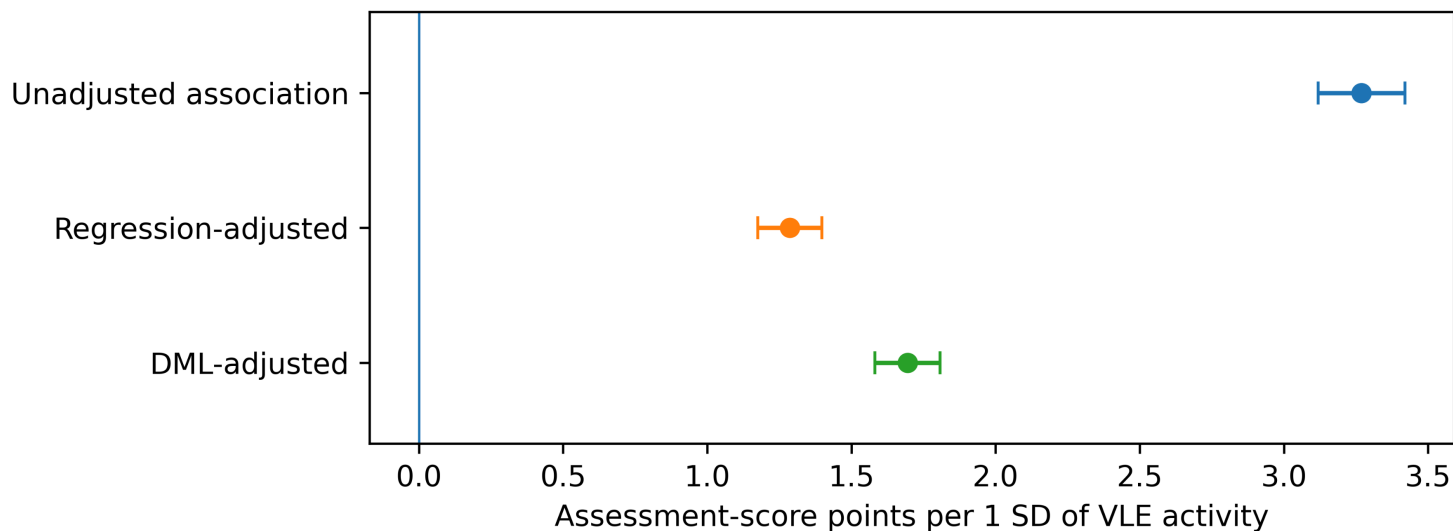


Figure 2

Unadjusted, Regression-Adjusted, and DML-Adjusted Associations Between VLE Activity and Assessment Performance

Note. Error bars indicate 95% confidence intervals. DML = double machine learning; VLE = virtual learning environment.

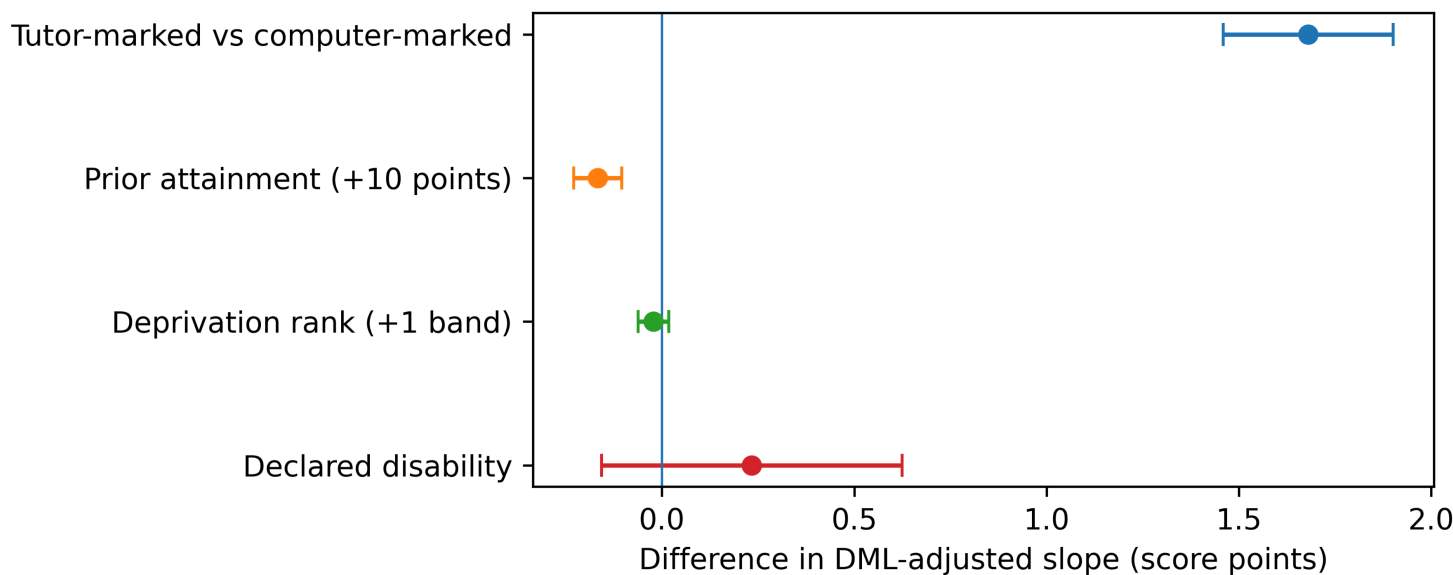


Figure 3

Prespecified Effect-Modifier Contrasts

Note. Points show differences in the DML-adjusted slope; error bars indicate 95% learner-clustered confidence intervals. DML = double machine learning.

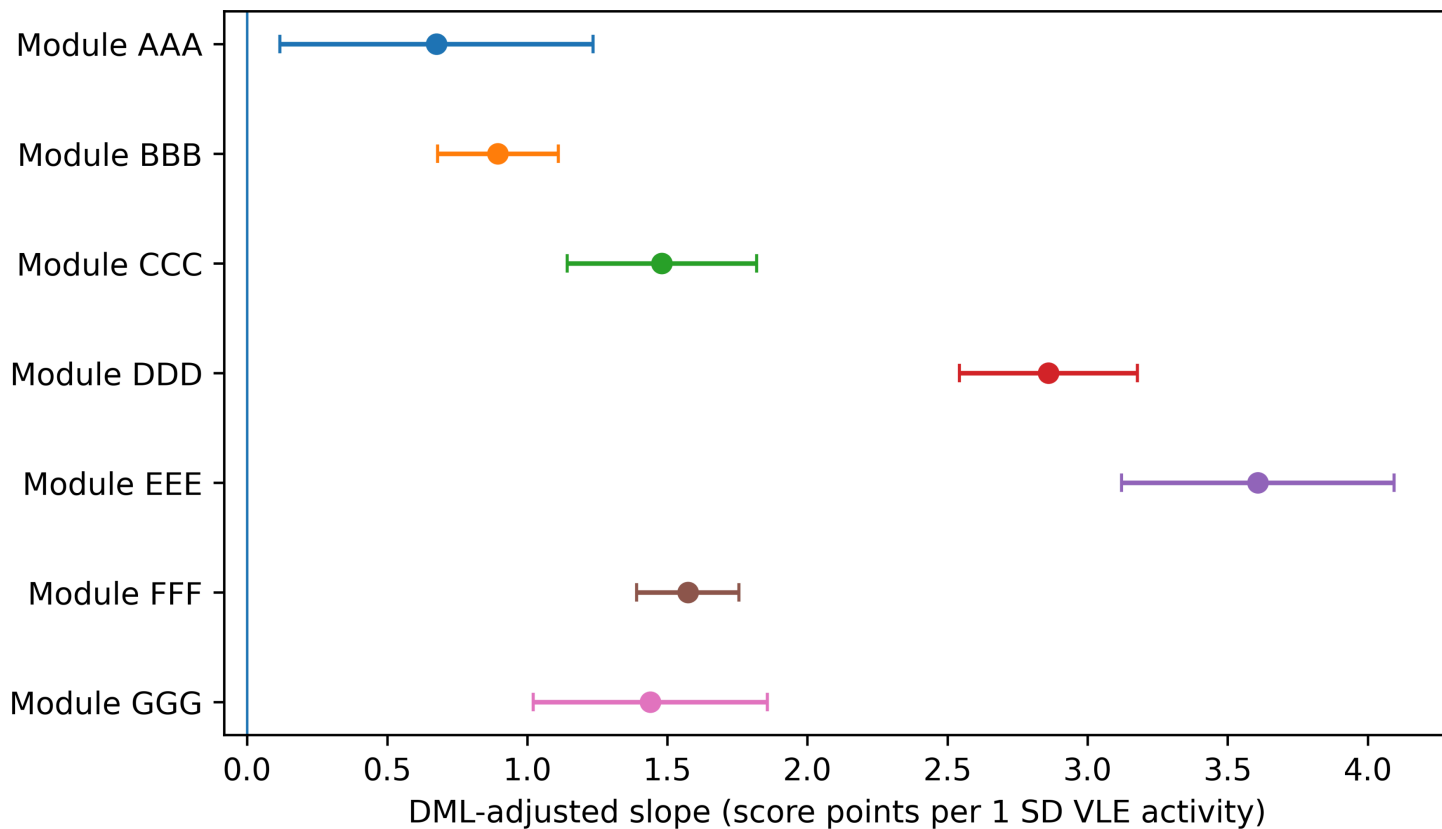


Figure 4

Module-Specific DML-Adjusted Slopes

Note. These are within-corpus sensitivity estimates rather than independent replications; error bars indicate 95% confidence intervals. DML = double machine learning.

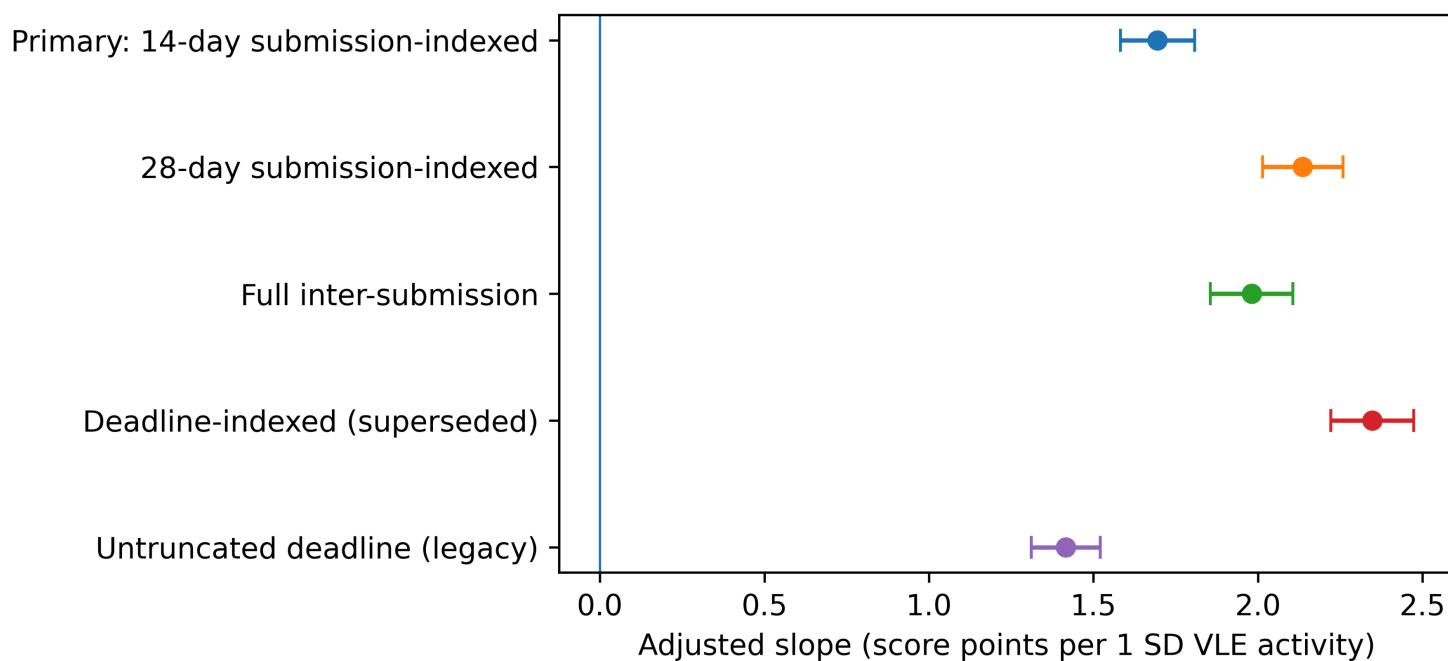


Figure 5

Sensitivity of the Adjusted Slope to Alternative Exposure-Window Definitions

Note. Earlier deadline-indexed definitions are retained as transparent sensitivity analyses; error bars indicate 95% confidence intervals. *SD* = standard deviation; VLE = virtual learning environment.