Higher Horizons Tracking Report 2026 Technical Annex: statistical analysis of progression rates of Higher Horizons participants and non-participants

Hannah Merry

May 2026

www.higherhorizons.co.uk

Contents

Table of Figures 4
Introduction 5
Constructing the Analysis Dataset 6
Restriction to 2019–2022 Expected Entry Cohorts 6
Exclusion of Mature Participants 6
Exclusion of Mature Non-Participants 7
Removal of Pre-Expected Entry Progressors 7
Exclusion of Students with Missing Demographic Data 7
Exploratory Data Tables 9
Contact Hours Distribution 9
Progression Rates by Contact Hours Band 9
Activity Types Across Cohorts 10
Logistic Regression – Full Results and Diagnostics 12
Model 1: Baseline 12
Model 2: Adding Demographic Controls 13
Multicollinearity Check 15
Linearity Check (Box–Tidwell Test) 15
Sensitivity Check: Free School Meals Eligibility 16
Sensitivity Check: Disability Status 17
Model 3: Dose-Response 17
Sensitivity Check: Band Cut-Points (Continuous Specifications) 19
Sensitivity Check: Year-Stratified Dose-Response 20
Model 4: Activity Types 20
Sensitivity Check: Contact Hours Replacing Binary Flags 21
Validation (Models 1–4 compared) 22
Propensity Score Methods – Full Results and Diagnostics 23
Participation Analysis 23
Propensity Score Model 23
Propensity Score Matching (PSM): Participation 23
Coarsened Exact Matching (CEM): Participation 24

Inverse Probability of Treatment Weighting (IPTW): Participation 25
Participation Analysis Summary 25
Dose-Response Propensity Score Analysis 26
Propensity Score Matching (PSM): Dose-Response 27
Coarsened Exact Matching (CEM): Dose-Response 28
Inverse Probability of Treatment Weighting (IPTW): Dose-Response 28
Dose-Response Analysis Summary 28

Table of Figures

Table 1: Summary of dataset filtering steps 6
Table 2: Missing data rates by variable and group 8
Table 3: Contact Hours Summary Statistics by Expected HE Entry Year 9
Table 4: Proportion of Participants in Each Contact Hour Band 9
Table 5: Progression Rate by Contact Hours Band and Expected HE Entry Year (all participants) 9
Table 6: Progression Rate by Contact Hours Band and Expected HE Entry Year (TUNDRA Q1 and Q2 students only) 10
Table 7: Mean number of each activity type per participant, by Expected HE Entry Year 10
Table 8: Proportion (%) of participants who attended at least one activity of each type, by Expected HE Entry Year 10
Table 9: Percentage of participants attending 1 / 2 / 3 / 4+ distinct activity types, by Expected HE Entry Year 10
Table 10: Logistic Regression Model 1 Results 12
Figure 1: Calibration Plot Model 1 13
Table 11: Logistic Regression Model 2 Results 13
Figure 2: Calibration Plot Model 2 15
Table 12: Results of Area-Based Sensitivity Check 15
Table 13: Results of FSM Sensitivity Check 16
Table 14: Results of FSM Categorical Sensitivity Check 16
Table 15: Results of Disability Sensitivity Check 17
Table 16: Logistic Regression Model 3 Results 18
Figure 3: Calibration Plot Model 3 18
Table 17: Sensitivity Check – Dose-Response Specifications 19
Table 18: Year-Stratified Dose-Response Odds Ratios (Model 3) 20
Table 19: Logistic Regression Model 4 Results 20
Figure 4: Calibration plot Model 4 21
Table 20: Validation Metrics for All Logistic Regression Models 22
Table 21: Summary of Results from All Methods 26
Table 22: Dose-Response Propensity Score Results 28

Introduction

This technical annex accompanies the Higher Horizons Tracking Report 2026, which is available here. It assumes familiarity with the methodology, dataset and findings described in that report, and provides the additional analytical detail that underpins those findings. The annex is intended for readers who wish to scrutinise or reproduce the analysis such as researchers, evaluators, and technical staff.

The annex covers four areas:

1. The full detail of dataset construction, including the rationale for each filtering decision and the analysis of dropped students.
2. Exploratory data tables showing the distribution of contact hours, progression rates by contact hours band, and activity type engagement across cohort years.
3. Full logistic regression results for all four models, including validation metrics, diagnostic checks, and sensitivity checks for free school meals eligibility, disability status, and the dose-response specification.
4. Full propensity score matching results for both the participation and dose-response analyses, including the matched-versus-dropped analysis, Rosenbaum bounds sensitivity checks, and IPTW weight diagnostics.

All tracking data used in this analysis was provided by HEAT. Thank you to the HEAT Service for their work in facilitating access to, and preparing, this dataset.

Please note: the HESA Standard Rounding Methodology has been applied to all numbers provided in this report:

1. Counts of people are rounded to the nearest multiple of 5
2. Percentages* are not published if they are fractions of a small group of people (fewer than 22.5)
3. Averages are not published if they are averages of a small group of people (7 or fewer)

Rounding and suppression are applied only to published descriptive counts and percentages. Statistical models were fitted on the underlying individual-level analysis dataset without rounding.

Constructing the Analysis Dataset

The raw dataset contains 44,875 unique student records with expected HE entry years spanning 2017 to 2025. The filtering steps below were applied in sequence to produce the final analysis dataset. Table 1 summarises the cumulative effect of each step.

Table 1: Summary of dataset filtering steps

Filtering step	Records removed	Cumulative total	Remaining
Raw dataset	—	—	44,875
Restriction to 2019–2022 cohorts (excl. pre-2019 and 2023)	16,940	16,940	27,935
Exclusion of mature participants (outreach after expected entry year)	325	17,265	27,610
Exclusion of mature non-participants (data collected after expected entry year)	2,920	20,185	24,690
Removal of pre-expected entry progressors	30	20,215	24,660
Exclusion of students with missing demographic data	1,875	22,090	22,790
Final analysis dataset	**—**	**—**	**22,790**

Restriction to 2019–2022 Expected Entry Cohorts

Cohorts with expected entry years before 2019 were removed for two reasons. Firstly, they contained very small sample sizes (between 30 and 385 students per year), which would produce statistically unstable estimates. Secondly, and more importantly, there were no non-participant records in these earlier years due to changes in data collection. Without a comparison group it is not possible to distinguish the effect of outreach participation from the baseline progression rate. This step removed 1,920 records.

The 2023 cohort was excluded because their by-19 tracking window was not yet complete at the time of analysis. Determining whether 2023 students entered HE by age 19 requires HESA data for the 2024/25 academic year, which was not available. This was by far the largest cohort; 15,020 records were removed. In total, 16,940 students were removed at this stage.

Exclusion of Mature Participants

Participants who engaged in outreach after their expected HE entry year were removed (325 students). These individuals had already passed the outcome window before engaging with the programme and could not have progressed to HE by age 19 as a result of outreach participation. They represent a distinct demographic group who had taken a

non-linear route through education and are not directly comparable with typical participants.

Exclusion of Mature Non-Participants

Non-participants whose records were collected after their expected HE entry year were removed (2,920 students). These individuals were still present at a Higher Horizons target school or college after the point at which they would have been expected to progress to HE, suggesting they were still in further education rather than being eligible for the by-19 outcome window.

Including them would have introduced a serious downward bias in the non-participant progression rate. These late-captured non-participants had a progression rate of just 8.7%, compared with 30.7% among non-participants whose data was collected within the normal timeframe. Without removal, they would have artificially depressed the non-participant progression rate, overstating the apparent effect of outreach participation.

This filter was not applied to participants, for whom activity records were used as the maturity filter instead. The date of data collection in HEAT is a manually entered administrative field that reflects when data was received by Higher Horizons or when the record was last updated/saved in the previous tracking service EMWPREP, not when the student actually participated. Among the 2,140 participants who would have been removed by this filter, 40.6% had progressed to HE, a rate only 1.37 percentage points higher than the retained participant group (39.2%). This narrow gap, combined with the implausibility of students being simultaneously in further education and recorded in the HESA student data, indicates that the data collection dates for these participants are unreliable. The activity-based filter already excludes participants who genuinely engaged after the outcome window.

Removal of Pre-Expected Entry Progressors

Thirty records showed students entering HE before their expected entry year. While early entry is possible in limited circumstances, the small number of such cases and the high likelihood that they represent data entry errors led to their removal.

Exclusion of Students with Missing Demographic Data

The logistic regression and propensity score matching methods both require complete demographic data for every student. Four variables are used as covariates: sex, ethnicity (grouped), TUNDRA quintile, and IMD quintile. Students with missing data on any of these four fields were excluded.

Table 2: Missing data rates by variable and group

Variable	Missing: participants	Missing: non-participants	Students excluded
Sex	2.69%	2.12%	—
Ethnicity	2.30%	6.42%	—
TUNDRA Quintile	0.93%	2.31%	—
IMD Quintile	0.06%	0.22%	—
Any of the above (combined exclusion)	**—**	**—**	**1,875**

The combined exclusion removed 1,875 records, slightly fewer than the sum of individual variable exclusions because some students were missing on more than one field. The dropped students are not a random subset of the dataset. They progressed to HE at lower rates than those retained: approximately 7 percentage points lower for non-participants and 10 percentage points lower for participants. Dropped participants were also more concentrated at IMD quintile 1, and nearly half had missing sex data. This group appears to represent some of the most disadvantaged students in the dataset, whose records are incomplete across multiple fields.

The practical consequence is a small downward bias in the estimated treatment effects. Before exclusion, the raw progression gap between participants and non-participants was 8.8 percentage points; after exclusion it narrowed to 8.5 percentage points. The regression and propensity score matching estimates presented in subsequent sections are therefore likely to be marginally conservative.

Exploratory Data Tables

This section provides detailed data tables supporting the exploratory analysis presented in the summary report. These tables are referenced in the summary but not reproduced in full there.

Contact Hours Distribution

Table 3: Contact Hours Summary Statistics by Expected HE Entry Year

Statistic	2019	2020	2021	2022
Mean	6.7	6.7	11.5	13.3
Standard Error	0.2	0.1	0.2	0.2
Median	6.0	5.5	7.0	8.0
Mode	6.0	1.0	2.0	1.0
Standard Deviation	7.4	6.5	12.9	14.5
Sample Variance	55.1	42.7	166.1	209.8
Kurtosis	10.2	14.1	7.1	6.8
Skewness	2.8	2.9	2.3	2.2
Range	62.5	65.6	133.7	135.3
Minimum	0.5	0.5	0.3	0.2
Maximum	63.0	66.0	134.0	135.5
Sum	14,753.6	16,805.5	45,522.3	64,747.1
Sample Size	**2,200**	**2,520**	**3,960**	**4,865**

Table 4: Proportion of Participants in Each Contact Hour Band

Year	Low (<3h)	Medium (≥3 and <8h)	High (≥8h)	Median (h)
2019	31.5%	47.4%	21.1%	6.0
2020	29.2%	45.3%	25.5%	5.5
2021	23.4%	35.5%	41.1%	7.0
2022	19.2%	32.3%	48.5%	8.0

Progression Rates by Contact Hours Band

Table 5: Progression Rate by Contact Hours Band and Expected HE Entry Year (all participants)

Year	Overall	Low (<3h)	Medium (≥3 and <8h)	High (≥8h)
2019	51.3%	41.8%	54.3%	58.7%
2020	43.3%	38.2%	43.9%	47.4%
2021	37.6%	35.2%	39.2%	37.7%
2022	35.4%	32.5%	36.6%	35.7%

Table 6: Progression Rate by Contact Hours Band and Expected HE Entry Year (TUNDRA Q1 and Q2 students only)

Year	Overall	Low (<3h)	Medium (≥3 and <8h)	High (≥8h)
2019	46.1%	30.8%	50.6%	55.4%
2020	39.0%	30.8%	38.2%	46.4%
2021	32.5%	25.5%	32.5%	35.0%
2022	31.0%	25.4%	30.6%	32.8%

The dose-response gradient is inconsistent in the full cohort in 2021 and 2022, where High and Medium rates converge. Restricting to TUNDRA Q1 and Q2 students reveals a cleaner and more consistent gradient across all four years, consistent with the interpretation that the flattening in the full cohort reflects compositional change rather than a genuine weakening of the dose-response relationship.

Activity Types Across Cohorts

Table 7: Mean number of each activity type per participant, by Expected HE Entry Year

Year	Campus Visit	Summer School	IAG	Attainment Raising	Subject Specific	Mentoring	Other
2019	0.44	0.11	0.92	0.13	0.10	0.07	0.01
2020	0.54	0.02	0.94	0.54	0.20	0.10	0.01
2021	0.56	0.21	1.82	0.45	0.37	0.22	0.03
2022	0.65	0.11	1.75	0.54	0.60	0.43	0.04

Table 8: Proportion (%) of participants who attended at least one activity of each type, by Expected HE Entry Year

Year	Campus Visit	Summer School	IAG	Attainment Raising	Subject Specific	Mentoring	Other
2019	39.9	3.4	53.2	10.9	8.3	2.6	1.1
2020	43.8	0.8	55.1	25.4	15.4	2.8	0.1
2021	40.0	7.0	76.5	26.7	16.5	2.8	0.5
2022	43.5	4.3	77.3	28.3	29.3	3.4	0.5

Table 9: Percentage of participants attending 1 / 2 / 3 / 4+ distinct activity types, by Expected HE Entry Year

Year	1 type	2 types	3 types	4+ types
2019	83.7%	13.7%	2.0%	0.6%
2020	64.6%	27.8%	7.1%	0.4%
2021	52.7%	29.9%	12.6%	4.8%
2022	48.2%	27.4%	15.3%	9.2%

The share of participants attending only one type of activity fell from 84% in 2019 to 48% in 2022. The share attending three or more types rose from 2.6% to 24.5% over the same period. This diversification reflects both the longer period over which later cohorts could engage with the programme and the expansion of the programme's activity offer.

Logistic Regression – Full Results and Diagnostics

Four logistic regression models were fitted in sequence. This section presents the full results tables, validation metrics, diagnostic checks and sensitivity checks for each model. For an explanation of the modelling approach and interpretation of the findings, see the summary report.

Four standard diagnostics are reported for each model. The AUC-ROC measures discrimination (how well the model distinguishes progressors from non-progressors), with 0.5 indicating chance and values of 0.7 or above considered acceptable. The Brier Score measures average prediction error, with lower scores indicating a better fit. The Hosmer–Lemeshow test assesses calibration by comparing observed and expected progression rates across deciles of predicted probability; a p-value above 0.05 indicates no evidence of poor calibration. The Pseudo R-squared indicates the share of variation explained. Because these are behavioural outcomes with many unobserved determinants, AUC values below conventional thresholds are expected; calibration is the more relevant diagnostic for the purpose of estimating a participation effect.

Model 1: Baseline

Model 1 estimates the association between outreach participation and progression by age 19, controlling only for expected HE entry year. Its purpose is to establish a baseline for comparison rather than to produce a final estimate.

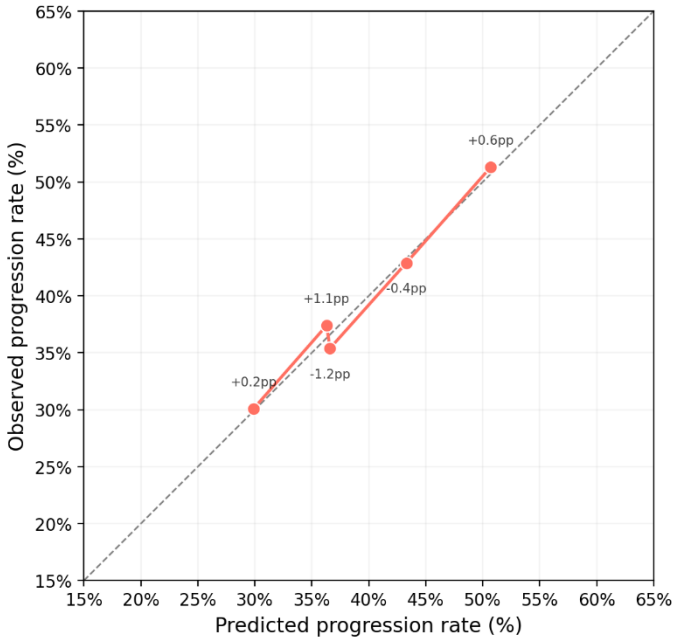
Table 10: Logistic Regression Model 1 Results

Term	Odds Ratio	95% Confidence Interval	p-value
Participated (vs not participated)	1.348	1.273 – 1.428	<0.001
Year 2020 (vs 2019)	0.743	0.673 – 0.821	<0.001
Year 2021 (vs 2019)	0.555	0.508 – 0.607	<0.001
Year 2022 (vs 2019)	0.562	0.516 – 0.612	<0.001

Validation: Four standard diagnostics were run to assess how well Model 1 fits the data. The AUC-ROC of 0.571 is weak, indicating that the model has limited ability to distinguish individual students who progressed from those who did not. This is expected given that the model contains only two predictors: participation status and cohort year. A value close to 0.5 indicates performance only marginally better than chance at the individual level. The Brier Score of 0.228 represents a small improvement over the naïve baseline of 0.232, which simply predicts the overall mean progression rate for every student. This confirms that adding participation and year as predictors does improve on that baseline, but only modestly. The Hosmer-Lemeshow test (chi-squared = 6.31, p = 0.098) provides no significant evidence of miscalibration, though the margin is narrower

than for the later models: when the model assigns a group of students a 40% probability of progressing, approximately 40% of them do. The maximum deviation between observed and expected progression rates across the five risk groups is 1.2 percentage points (Figure 1). The pseudo R-squared of 0.013 reflects the limited explanatory power of the two-predictor specification.

Figure 1: Calibration Plot Model 1



Taken together, these diagnostics tell a coherent story: Model 1 is a poor tool for predicting which individual student will progress, but its group-level probability estimates are reliable. Because the model's purpose is to establish a baseline participation estimate rather than to predict individual outcomes, calibration is the more relevant property, and on that measure the model performs well.

Model 2: Adding Demographic Controls

Model 2 adds sex, ethnicity (grouped), TUNDRA quintile and IMD quintile to the Model 1 specification. This is the primary model for estimating the adjusted participation effect.

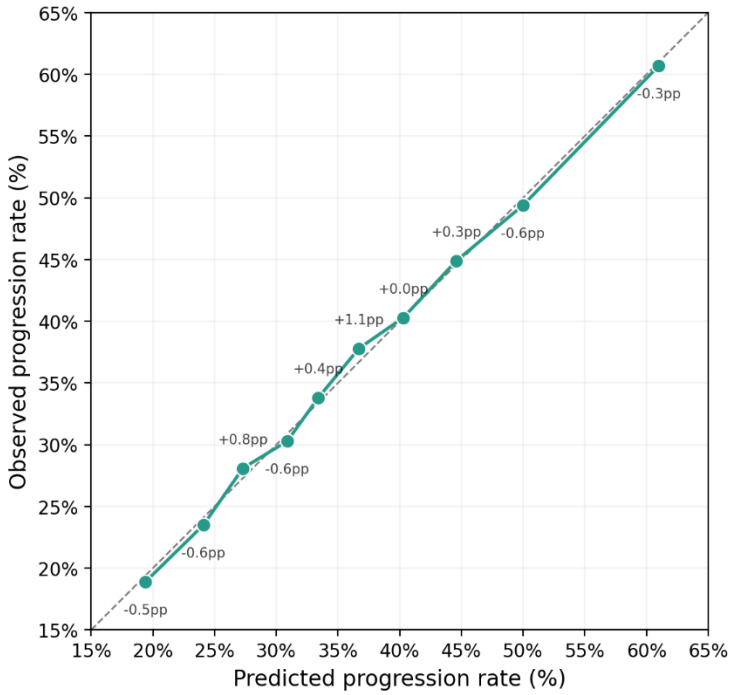
Table 11: Logistic Regression Model 2 Results

Term	Odds Ratio	95% Confidence Interval	p-value
Participated (vs not participated)	1.519	1.430 – 1.614	<0.001
Year 2020 (vs 2019)	0.800	0.723 – 0.886	<0.001
Year 2021 (vs 2019)	0.599	0.546 – 0.657	<0.001
Year 2022 (vs 2019)	0.626	0.574 – 0.684	<0.001
Sex (Male vs Female)	0.564	0.533 – 0.596	<0.001
Ethnicity – Arab (vs White)	0.978	0.536 – 1.784	0.943

Ethnicity – Asian (vs White)	2.662	2.372 – 2.988	<0.001
Ethnicity – Black (vs White)	1.629	1.318 – 2.013	<0.001
Ethnicity – Mixed (vs White)	1.361	1.180 – 1.568	<0.001
Ethnicity – Other (vs White)	1.055	0.658 – 1.693	0.823
TUNDRA Quintile (per quintile higher)	1.115	1.082 – 1.149	<0.001
IMD Quintile (per quintile higher)	1.157	1.124 – 1.191	<0.001

Validation: Model 2 shows clear improvements across all diagnostic metrics relative to Model 1. The AUC-ROC rises from 0.571 to 0.649, an improvement of 0.077. This is a meaningful gain and reflects the genuine predictive information carried by the five demographic variables added to the model. The value remains below the conventional threshold of 0.7 for acceptable discrimination, which is to be expected for behavioural outcomes like HE progression that depend on a wide range of factors not captured in this dataset, most notably prior attainment. The Brier Score improves from 0.228 to 0.217, confirming that Model 2 makes better individual-level predictions than the baseline specification. The pseudo R-squared rises from 0.013 to 0.050, a near four-fold increase, though the model still explains only a small fraction of the total variation in outcomes. The Hosmer-Lemeshow test (chi-squared = 3.43, p = 0.905) confirms that calibration remains good, with a smaller chi-squared statistic than Model 1, indicating the demographic variables are genuinely improving the model's ability to assign accurate progression probabilities. The maximum deviation between observed and expected progression rates across decile groups is 1.1 percentage points (Figure 2). At the Youden's J optimal threshold of 0.365, the model correctly identifies 59% of true progressors and 63% of true non-progressors, an improvement on Model 1's sensitivity of 30% and a substantially better balance between the two. Overall accuracy is 61%.

Figure 2: Calibration Plot Model 2



Multicollinearity Check

Variance Inflation Factors (VIF) were calculated for all predictors. TUNDRA Quintile (VIF 9.06) and IMD Quintile (VIF 9.30) exceeded the standard investigation threshold of 5.0, reflecting a Pearson correlation of 0.715 between them. To test whether this distorts the participation estimate, Model 2 was re-run three times: with both TUNDRA and IMD included, with TUNDRA only, and with IMD only.

Table 12: Results of Area-Based Sensitivity Check

Version	Participation OR
Full Model 2 (both TUNDRA and IMD)	1.519
TUNDRA only	1.476
IMD only	1.493

The participation effect is essentially identical regardless of which area-based measure is included (range: 1.476 to 1.519). The multicollinearity does not distort the headline finding and there is no reason to drop either variable. Individual TUNDRA and IMD coefficients should not, however, be over-interpreted as fully independent effects.

Linearity Check (Box–Tidwell Test)

The Box–Tidwell test was applied to TUNDRA and IMD quintile to verify that each variable has a linear relationship with the log-odds of progression. For TUNDRA Quintile, p = 0.359: the linearity assumption holds. For IMD Quintile, p = 0.003,

indicating a slight non-linearity. Model 2 was re-run with both variables treated as categorical (five separate indicators per variable) to test whether this affects the participation estimate. The participation odds ratio under this specification is 1.526, compared to 1.519 in the primary model (difference: 0.007). The non-linearity in IMD does not affect the headline finding.

Sensitivity Check: Free School Meals Eligibility

FSM eligibility is an individual-level measure of economic disadvantage not included in the primary model due to missing data (11.1% of participants and 30.6% of non-participants). To test whether excluding FSM changes the headline finding, Models 1 and 2 were re-run on the 18,465 students for whom FSM data is available, with FSM added as a binary predictor.

Table 13: Results of FSM Sensitivity Check

Metric	Primary analysis (n = 22,790)	FSM sensitivity (n = 18,465)
Model 1 participation OR	1.348	1.732
Model 2 participation OR	1.519	1.948
Raw participant progression rate	40.1%	40.7%
Raw non-participant progression rate	31.6%	27.0%
Raw gap (percentage points)	8.5	13.7

The participation odds ratio rises from 1.519 to 1.948 in the restricted sample. However, this increase is already present in Model 1 (from 1.348 to 1.732) before FSM is added, which means it is driven by sample restriction rather than the FSM variable. The non-participants excluded by the FSM restriction progress at substantially higher rates than those retained (confirmed by the fall in raw non-participant progression from 31.6% to 27.0%), creating a biased comparison. The 1.948 figure overstates the participation effect.

A further test retained all 22,790 students by coding FSM as a three-category variable (Yes / No / Unknown).

Table 14: Results of FSM Categorical Sensitivity Check

Metric	Primary Model 2 (no FSM)	FSM-categorical Model 2
Sample size	22,790	22,790
Participation OR	1.519	1.631
FSM "Yes" OR (vs "No")	—	0.626 ($p < 0.001$)
FSM "Unknown" OR (vs "No")	—	1.339 ($p < 0.001$)

AUC	0.649	0.661
Pseudo R-squared	0.050	0.057
Hosmer–Lemeshow p-value	0.905	0.033

The participation odds ratio with this specification is 1.631, smaller than the binary approach (1.948) because it avoids sample-loss bias. The Unknown category has an odds ratio of 1.339 ($p < 0.001$), confirming that students with missing FSM data are systematically more likely to progress to HE and are not a random subset of the dataset. Because the FSM categorical model had worse validation than the primary Model 2 (Hosmer–Lemeshow p = 0.033 versus 0.905) and did not change the direction or significance of the participation finding, FSM eligibility was not added to the primary model.

Sensitivity Check: Disability Status

Disability data was missing for 8.6% of the analysis sample. Models 1 and 2 were re-run on the 20,830 students for whom disability data is available, with disability added as a binary predictor.

Table 15: Results of Disability Sensitivity Check

Metric	Primary analysis (n = 22,790)	Disability sensitivity (n = 20,830)
Model 1 participation OR	1.348	1.416
Model 2 participation OR	1.519	1.525
Raw participant progression rate	40.1%	40.2%
Raw non-participant progression rate	31.6%	30.4%
Raw gap (percentage points)	8.5	9.8

The participation odds ratio in Model 2 is essentially unchanged: 1.525 versus 1.519 in the primary model. The disability variable itself is a strong predictor (OR 0.548, $p < 0.001$), and adding it marginally improves diagnostic metrics (AUC 0.663 versus 0.649; Brier 0.214 versus 0.217). VIF for disability is 1.25. However, because the participation estimate is identical whether disability is included or not, and because including it requires losing 1,960 students, disability was not added to the primary model. This is an analytical choice rather than a data quality constraint.

Model 3: Dose-Response

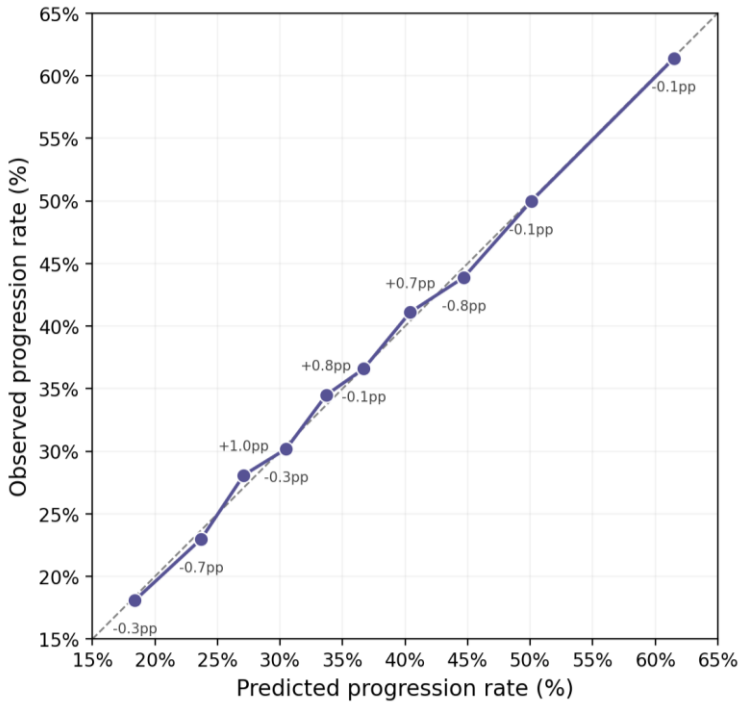
Model 3 replaces the binary participation indicator with a four-category contact hours measure: No contact (reference), Low (<3 hours), Medium (≥3 and <8 hours), High (≥8 hours).

Table 16: Logistic Regression Model 3 Results

Contact hours band	Odds Ratio	95% Confidence Interval	p-value
Low (<3h) vs No contact	1.171	1.07 – 1.28	<0.001
Medium (≥3 and <8h) vs No contact	1.574	1.46 – 1.70	<0.001
High (≥8h) vs No contact	1.735	1.61 – 1.87	<0.001

Validation: Model 3 produces small but consistent improvements over Model 2 across the discrimination and predictive accuracy metrics. The AUC rises from 0.649 to 0.652 (+0.003) and the Brier Score is unchanged at 0.217. These modest gains reflect the nature of the change between the two models: Model 2 added five new demographic variables containing genuinely new information about each student, whereas Model 3 modifies an existing variable by splitting the binary participation flag into a four-level contact hours scale. No new information about students is introduced; the improvement comes from structuring the existing engagement data more precisely. That the AUC increases at all confirms that the contact hours bands carry more predictive signal than the binary participation flag alone. The Hosmer-Lemeshow test (chi-squared = 3.920, p = 0.864) indicates good calibration, comparable to Model 2 and well above the 0.05 threshold. The pseudo R-squared rises marginally to 0.052. At the Youden's J optimal threshold of 0.375, sensitivity is 58% and specificity is 64%, broadly similar to Model 2. Overall accuracy is 62%. The maximum deviation between observed and expected progression rates across decile groups is 1.0 percentage points (Figure 3).

Figure 3: Calibration Plot Model 3



The diagnostics confirm that replacing the binary participation indicator with a banded contact hours measure does not compromise model fit, and produces a marginally better-fitting specification that allows the participation effect to be disaggregated by engagement intensity.

Sensitivity Check: Band Cut-Points (Continuous Specifications)

To test whether the dose-response gradient is an artefact of the chosen band cut-points, Model 3 was re-fitted using contact hours as a continuous predictor (log-transformed and linear) on the participant-only sample (n = 13,545).

Table 17: Sensitivity Check – Dose-Response Specifications

Specification	Dose-response measure	Effect	95% CI	p-value
Banded (full sample)	Low (<3h) vs No contact	OR 1.171	1.07–1.28	<0.001
Banded (full sample)	Medium (≥3 and <8h) vs No contact	OR 1.574	1.46–1.70	<0.001
Banded (full sample)	High (≥8h) vs No contact	OR 1.735	1.61–1.87	<0.001
Banded (participants only)	Medium vs Low	OR 1.362	1.24–1.50	<0.001
Banded (participants only)	High vs Low	OR 1.545	1.40–1.70	<0.001
Log continuous (participants only)	Per doubling of contact hours	OR 1.170	1.14–1.20	<0.001
Linear continuous (participants only)	Per additional 10 hours	OR 1.155	1.12–1.19	<0.001

All three specifications produce positive, statistically significant dose-response effects of comparable magnitude. Model fit is virtually identical across specifications (AUC 0.656–0.658; Brier 0.222–0.223; both pass Hosmer–Lemeshow). The log specification fits better than the linear, consistent with diminishing returns: each additional hour has its largest effect at low engagement levels and progressively smaller effects at higher levels. The dose-response finding is not driven by the band cut-points.

Sensitivity Check: Year-Stratified Dose-Response

Model 3 was re-fitted separately within each cohort year to test whether the dose-response gradient holds within years rather than being driven by between-cohort compositional change.

Table 18: Year-Stratified Dose-Response Odds Ratios (Model 3)

Contact hours band	2019	2020	2021	2022
Low (<3h) vs No contact	1.08 (0.87–1.34)	1.20 (0.98–1.47)	1.45 (1.22–1.71)*	1.12 (0.96–1.31)
Medium (≥3 and <8h) vs No contact	1.81 (1.48–2.21)*	1.51 (1.25–1.82)*	1.92 (1.65–2.23)*	1.39 (1.23–1.57)*
High (≥8h) vs No contact	2.41 (1.89–3.08)*	1.87 (1.52–2.30)*	1.99 (1.73–2.30)*	1.49 (1.34–1.67)*

* $p < 0.001$

The Low < Medium < High ordering holds in every cohort year after demographic adjustment. Medium and High are statistically significant in all four years. The Low band is not statistically significant in three of four years (2019, 2020, 2022), suggesting that under 3 hours of engagement may not consistently produce a detectable association with progression. The diminishing-returns pattern is confirmed within each year: the step from Medium to High is smaller than the step from Low/No-contact to Medium in every cohort. High-band odds ratios range from 1.49 in 2022 to 2.41 in 2019, consistent with the year-on-year variation in the overall participation effect documented in the year-stratified Model 2.

Model 4: Activity Types

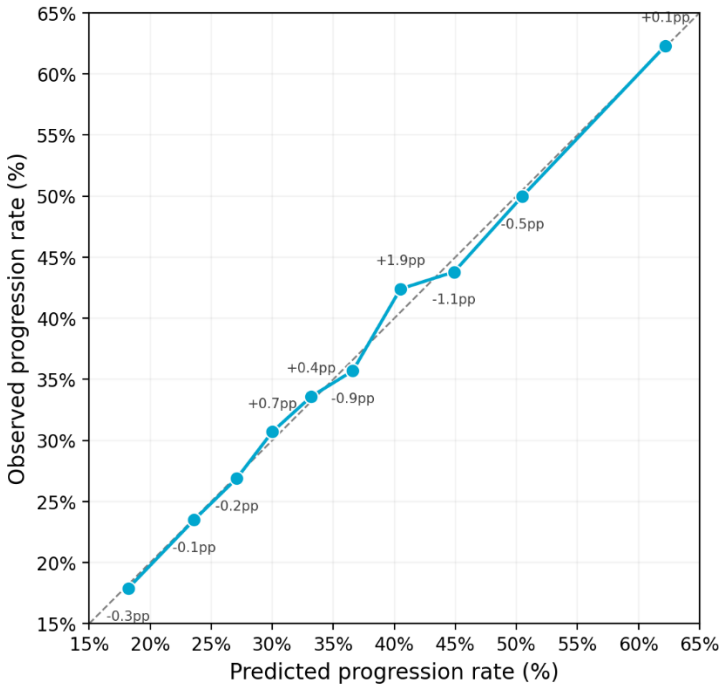
Table 19: Logistic Regression Model 4 Results

Activity type	N attended	% of sample	Odds Ratio	95% Confidence Interval	p-value
Campus Visit	5,680	24.9%	1.593	1.49 – 1.70	<0.001
Other Type	70	0.3%	1.558	0.95 – 2.56	0.081
Summer School	585	2.6%	1.353	1.14 – 1.61	<0.001
IAG	9,350	41.0%	1.258	1.19 – 1.34	<0.001
Mentoring Programme	405	1.8%	1.256	1.02 – 1.55	0.032
Subject Specific	2,655	11.6%	1.066	0.97 – 1.17	0.168
Attainment Raising	3,315	14.5%	0.850	0.78 – 0.93	<0.001

Validation: Model 4 is statistically and practically comparable to Models 2 and 3 across all diagnostic metrics. The AUC of 0.655 is fractionally higher than Model 3's 0.652 and

Model 2's 0.649. The Brier Score falls marginally to 0.216. The pseudo R-squared rises to 0.054. These are very small improvements in absolute terms, reflecting the fact that all three models contain broadly similar information; the differences between them lie in how participation is structured rather than in the volume of information available. The Hosmer-Lemeshow test (p = 0.595) confirms reasonable calibration, with a maximum deviation of 1.9 percentage points between observed and expected progression rates across decile groups (Figure 4). This is marginally higher than in Models 2 and 3 but remains well within an acceptable range for estimating a participation effect. At the Youden's J optimal threshold, sensitivity is 59% and specificity is 64%. A multicollinearity check found Pearson correlations between the seven activity-type flags to be small throughout, with the largest positive correlation of 0.148 between Attainment Raising and Subject Specific, and the most negative of -0.285 between Campus Visit and IAG. Variance Inflation Factors for all activity-type flags fall between 1.0 and 1.15, well below the 5.0 threshold for investigation. The activity-type coefficients are therefore independently identified and the differences between them reflect genuine variation in their association with progression rather than statistical artefacts of correlated predictors.

Figure 4: Calibration plot Model 4



Sensitivity Check: Contact Hours Replacing Binary Flags

Model 4 was re-fitted twice with the seven binary flags replaced by (a) raw contact hours per activity type and (b) log-transformed contact hours per activity type. The pattern of results is consistent across all three specifications: Campus Visit remains the strongest positive predictor; IAG and Summer Schools are positive and significant; Attainment

Raising is consistently the weakest-performing type; Subject Specific is not significant in any version. Model fit is essentially identical across specifications (AUC 0.653–0.655; Brier 0.216; all pass Hosmer–Lemeshow). The activity-type findings are not an artefact of how engagement is measured.

The negative coefficient (OR 0.85) for attainment raising activities most likely reflects teacher selection of lower-attaining students into these activities rather than a genuine negative effect of the activities themselves. In the year-stratified sensitivity check, the odds ratio for Attainment Raising rose from 0.646 in 2019 to 1.030 in 2022, but was not statistically significant in 2021 ($p = 0.54$) or 2022 ($p = 0.69$). Without prior attainment data these results cannot be fully interpreted.

Validation (Models 1–4 compared)

Table 20: Validation Metrics for All Logistic Regression Models

Metric	Model 1	Model 2	Model 3	Model 4
AUC-ROC	0.571	0.649	0.652	0.655
Brier Score	0.228	0.217	0.217	0.216
Hosmer–Lemeshow p-value	0.098	0.905	0.864	0.595
Pseudo R-squared	0.013	0.050	0.052	0.054
Sensitivity at optimal threshold	0.301	0.588	0.576	0.588
Specificity at optimal threshold	0.798	0.627	0.644	0.639

Propensity Score Methods – Full Results and Diagnostics

This section provides the full analytical detail for the propensity score analyses summarised in the summary report. For the participation analysis, this includes the matched-versus-dropped analysis, balance diagnostics, bootstrapped confidence intervals and Rosenbaum bounds sensitivity checks. For the dose-response analysis, it includes the same information plus sub-group results for Medium and High-band students.

Participation Analysis

Propensity Score Model

The propensity score for each student was estimated using a logistic regression predicting participation from sex, ethnicity, TUNDRA quintile and IMD quintile, fitted separately within each cohort year to eliminate year-level confounding by construction. Propensity score model AUCs ranged from 0.574 in 2019 to 0.658 in 2022, indicating that the demographic variables capture meaningful but not deterministic selection into participation.

Propensity Score Matching (PSM): Participation

Specification: 1:1 nearest-neighbour matching without replacement, exact matching on cohort year, caliper of 0.2 standard deviations of the propensity score distribution.

Match rates: 7,720 pairs from 13,545 participants (57.0%). By year: 32.1% (2019), 37.9% (2020), 56.7% (2021), 78.4% (2022). The low rate in 2019 and 2020 reflects the limited non-participant control pool (705 and 970 students respectively).

Balance: Post-matching balance was assessed using standardised mean differences (SMDs) and variance ratios for each covariate. SMDs were well below the conventional threshold of 0.1 for sex, TUNDRA quintile, IMD quintile, and most ethnicity categories, and variance ratios fell within the target range of 0.8 to 1.25 across the same variables. The one exception was the Arab ethnicity group, where the variance ratio of 1.61 exceeded the target. This reflects the very small number of Arab students in the dataset (n = 50) rather than a systematic matching failure, and does not materially affect the overall estimate.

Result: +8.9 percentage points (95% CI: +7.4pp to +10.4pp, $p < 0.001$ using McNemar's test to estimate paired ATT). Bootstrapped confidence intervals (100 iterations): +7.9pp to +10.9pp. Effect positive and statistically significant in every cohort year (range: +6.3pp in 2022 to +12.7pp in 2021).

Matched vs Dropped Analysis

The 5,825 dropped participants (43.0%) differ systematically from the 7,720 matched. Dropped participants had a mean TUNDRA quintile of 1.84 versus 2.52 for matched participants, and a mean IMD quintile of 1.90 versus 2.84. They were more likely to be from Asian (11.4% vs 5.2%) and Black (2.6% vs 1.4%) ethnic backgrounds. In aggregate, dropped participants progressed at 39.8% compared to 40.3% for matched participants (a gap that is not statistically significant, p = 0.54). However, within each cohort year, matched participants consistently progressed at higher rates (+0.9pp in 2019, +3.5pp in 2020, +6.0pp in 2021, +7.2pp in 2022). The widening year-on-year gap reflects the improving control pool: as more non-participants become available in later cohorts, the matching becomes more selective. The PSM ATT applies primarily to the matched subset and may not fully represent the programme's most disadvantaged participants.

Sensitivity to Unobserved Confounding

Critical Γ = 1.45. An unobserved confounder would need to make a student 1.45 times more likely to participate, holding all observed demographics constant, to reduce the ATT to non-significance. This is a moderate level of robustness. Prior attainment is the most plausible omitted variable in this context. This finding is consistent with the interpretation that the results are associations rather than confirmed causal effects.

Based on a matched risk ratio of 1.28, an e-value of 1.89 was calculated. This means an unmeasured confounder would need to be associated with both participation and progression by a risk ratio of at least 1.89 each, above and beyond the measured covariates, to explain away the estimate. The corresponding value to move the confidence interval to include the null is 1.77. Prior attainment is still a plausible omitted variable in this context, but it should be noted participants are not chosen for Higher Horizons activities based on their attainment.

Coarsened Exact Matching (CEM): Participation

CEM matches participants to non-participants who share exactly the same profile across all variables: cohort year, sex, ethnicity, TUNDRA quintile and IMD quintile. This creates 678 unique demographic strata, of which 398 contain at least one participant and one non-participant. CEM uses variable match ratios: where a stratum has 30 participants and 3 non-participants, all 33 students are retained and each non-participant is weighted by 10. This avoids the control-pool drop problem that affects PSM.

Match rate: 13,125 of 13,545 participants (96.9%). Only 415 participants (3.1%) were dropped, compared to 43.0% under PSM. Match rate exceeded 94% in every cohort year. The 415 dropped students fell across 206 small strata representing sparse

demographic combinations where no non-participant with that exact profile existed in that year.

Balance: Because CEM achieves balance by construction rather than approximately, no post-matching balance check is required in the conventional sense: every standardised mean difference is exactly zero and every variance ratio exactly 1.0. This eliminates the balance uncertainty that affects PSM and removes the need for iterative balance checking across covariate specifications.

Result: +8.8 percentage points (95% CI: +7.3pp to +10.3pp, $p < 0.001$). Effect positive in every cohort year (range: +6.5pp in 2022 to +11.4pp in 2021). Within 0.1 percentage points of the PSM estimate despite the two methods using different matching approaches and retaining different proportions of participants.

Inverse Probability of Treatment Weighting (IPTW): Participation

IPTW reweights the non-participant group using the propensity score so that its weighted demographic distribution matches that of the participant group. No student is excluded.

Weight diagnostics: Median control weight 1.17; 99th percentile 4.57; maximum weight 12.37; only 7 of 9,245 controls had weights above 10; effective sample size 5,970 (64.6% of actual control count). ATT was stable across raw, trimmed and stabilised weight variants (+8.7pp for raw and stabilised; +8.8pp for trimmed).

Balance: Post-weighting balance was good on all covariates except the Other ethnicity category, where the variance ratio of 2.25 exceeded the 0.8 to 1.25 target range despite the SMD being small at 0.048. As with the Arab category in PSM, this reflects the very small size of this group ($n = 80$, 0.3% of the sample) rather than a substantive balance failure.

Result: +8.7 percentage points (95% CI: +7.2pp to +10.1pp, $p < 0.001$). Bootstrapped confidence intervals (500 iterations): +7.4pp to +10.2pp. Effect positive in every cohort year (range: +5.5pp in 2022 to +11.1pp in 2021).

Participation Analysis Summary

The table below summarised the results of all participation analyses. Because odds ratios are not directly comparable to the percentage-point treatment effects produced by the matching and weighting methods, the logistic regression estimate is additionally expressed as an average marginal effect via marginal standardisation (g-computation): each participant's predicted probability of progression was computed under participation and non-participation with their own covariates held fixed, and the differences averaged over the participant population, yielding the regression-based analogue of the ATT.

Table 21: Summary of Results from All Methods

Method	ATT / OR	95% CI	Sample size	Drop rate
Logistic regression (Model 2)	OR = 1.52	1.43–1.61	22,790	0%
Logistic regression (Model 2) marginal effects (g-computation)	+9.0pp	+7.7pp to +10.2pp	22,790	0%
Propensity score matching	+8.9pp	+7.4pp to +10.4pp	7,720 pairs	43.0%
Coarsened exact matching	+8.8pp	+7.3pp to +10.3pp	13,125 treated	3.1%
Inverse probability weighting	+8.7pp	+7.2pp to +10.1pp	22,790	0%

The three ATT-based methods produce estimates within a range of 0.2 percentage points of each other (+8.7pp to +8.9pp), with overlapping confidence intervals excluding zero. The direction of the gradient across methods (PSM > CEM > IPTW) is consistent with the matched-versus-dropped analysis: PSM excludes the most disadvantaged participants (slight upward bias), CEM retains nearly everyone (marginal downward adjustment), and IPTW includes everyone with no exclusions (most inclusive estimate). The magnitude of the gradient confirms that the bias introduced by exclusions is negligible. Expressed as an average marginal effect, the logistic regression model 2 estimate was +9.0 percentage points (95% CI: +7.7 to +10.2), placing the regression result within 0.3 percentage points of the matching and weighting estimates and confirming that the finding is not an artefact of any single analytical approach. The convergence across methods with very different assumptions, balance mechanisms and inclusion criteria provides strong evidence that the participation-progression association is genuine.

Dose-Response Propensity Score Analysis

Because Low-band participants (<3 hours) were largely indistinguishable from non-participants in the year-stratified Model 3 (significant in only 1 of 4 years), the Low band and non-participants were combined into a single control group for this analysis. The treated group comprises Medium and High-band participants (at least 3 hours of contact). This tests whether substantive engagement is associated with higher progression than minimal or no engagement.

The enlarged control pool (12,530 students: 9,245 non-participants and 3,290 Low-band participants) substantially reduces the control-pool imbalance that limited the participation PSM.

It should be noted that students do not self-select into more engagement with the programme, with the exception of Summer Schools. Schools book and arrange activities for their students. While teachers may influence which students are selected, it is not accurate to assume that higher engagement simply reflects greater student motivation.

Propensity Score Matching (PSM): Dose-Response

Match rate: 8,695 of 10,255 treated students (84.8%). Improvement from 57.0% in the participation analysis reflects the larger control pool. Year-specific match rates: 83.4% (2019), 77.6% (2020), 86.3% (2021), 91.6% (2022).

Balance: Post-matching balance was assessed using SMDs and variance ratios for each covariate. Unlike the participation analysis, balance was good on all covariates including all ethnicity categories: all SMDs were well below 0.1 and all variance ratios fell within the 0.8 to 1.25 target range. The absence of any balance failures represents an improvement on the participation PSM and reflects the better overlap in propensity score distributions produced by the larger control pool.

Result: +9.4 percentage points (95% CI: +8.1pp to +10.8pp, $p < 0.001$ using McNemar's test to estimate paired ATT and significance). Effect positive in every cohort year (range: +6.4pp in 2022 to +14.1pp in 2019). Bootstrapped confidence intervals (1000 iterations) were close to the analytic intervals (+8.4pp to +11.4pp), confirming that the analytic standard errors are adequate.

Matched vs dropped: 1,560 treated students dropped (15.2%). Dropped students were more disadvantaged (mean TUNDRA quintile 1.41 vs 2.25; mean IMD quintile 1.34 vs 2.51) and more concentrated in ethnic minority groups. Aggregate progression difference between matched and dropped was not statistically significant ($p = 0.55$). Within-year progression showed a more nuanced pattern: in 2019 and 2020, dropped students progressed slightly higher than matched; in 2021 and 2022 the pattern reversed. This likely reflects cohort composition effects rather than systematic bias in one direction.

Sub-group analysis: ATT estimated separately for Medium-band (+8.3pp, 4,115 pairs) and High-band (+10.4pp, 4,580 pairs) within the matched sample, consistent with the dose-response gradient from Model 3.

Sensitivity to omitted variables: Critical $\Gamma = 1.50$, slightly higher than the participation analysis (1.45). An unobserved confounder would need to increase the odds of substantive engagement by 1.50 times to reduce the dose-response ATT to non-significance. This modest improvement in robustness is consistent with a genuine dose-response effect being harder to explain away than a flat participation effect. The e-value also marginally improved to 1.91 from a matched risk ratio of 1.30.

Coarsened Exact Matching (CEM): Dose-Response

Match rate: 10,105 of 10,255 treated students (98.5%). Only 150 treated students were dropped (1.5%, compared to 3.1% in the participation analysis), spread across 104 small strata predominantly representing ethnic minority students in sparse demographic combinations. The improvement in match rate reflects the larger control pool available for the dose-response analysis: with Low-band participants now available as controls alongside non-participants, fewer demographic strata exist in which no control can be found.

Weight diagnostics: Median control weight 0.69; 99th percentile 2.20; no weights above 10; effective sample size 8,300 (67.6% of 12,280 controls used).

Result: +10.2 percentage points (95% CI: +8.8pp to +11.5pp, $p < 0.001$). Effect positive in every cohort year (range: +7.1pp in 2022 to +14.9pp in 2019). Balance perfect by construction.

Inverse Probability of Treatment Weighting (IPTW): Dose-Response

Weight diagnostics: The median control weight was 0.74, the 99th percentile was 1.88, and the maximum weight was 3.87. No controls had weights above 10, in contrast to seven controls in the participation analysis. The effective sample size of 9,430 represents 75.2% of the 12,530 controls, compared to 64.6% in the participation analysis, reflecting the better overlap in propensity score distributions when Low-band participants are available as controls alongside non-participants.

Balance: Post-weighting balance passed on all covariates across all three weight variants (raw, trimmed and stabilised), including the Other ethnicity category that showed a variance ratio above the target threshold in the participation analysis. This represents a clean balance result with no exceptions.

Result: +10.3 percentage points (95% CI: +9.0pp to +11.6pp, $p < 0.001$). Bootstrapped confidence intervals (487 iterations): +9.1pp to +11.6pp. Effect positive in every cohort year (range: +6.7pp in 2022 to +16.1pp in 2019).

Dose-Response Analysis Summary

Table 22: Dose-Response Propensity Score Results

Method	ATT	95% CI	Sample size	Drop rate
Propensity score matching	+9.4pp	+8.1pp to +10.8pp	8,695 pairs	15.2%
Coarsened exact matching	+10.2pp	+8.8pp to +11.5pp	10,105 treated	1.5%
Inverse probability weighting	+10.3pp	+9.0pp to +11.6pp	22,790	0%

The three methods produce estimates within a range of 0.9 percentage points (+9.4pp to +10.3pp), with all confidence intervals excluding zero. The direction of the gradient across methods is consistent with the matched-versus-dropped findings. The dose-response ATTs are consistently higher than the corresponding participation ATTs (+8.7pp to +8.9pp), indicating that substantive engagement (≥3 hours) is associated with a larger difference in progression rates than any participation at all. The improvement in analytical quality is also notable: match rates increased substantially across all methods, balance was clean with no covariate failures, and IPTW weight efficiency improved from 64.6% to 75.2%. The dose-response estimates are therefore not only larger than the participation estimates but also better supported by the underlying diagnostics.

Caveats: All three methods share the same vulnerability to unobserved confounders as the participation analysis. The Rosenbaum bounds critical Γ of 1.50 is moderate and so is the e-value of 1.91. Prior attainment remains the most likely uncontrolled variable. These findings should be interpreted as robust associations rather than confirmed causal effects.