Mathematics Support Use and Classroom-Relative Performance in First-Year University Mathematics

Heriberto Espino-Montelongo
Universidad de las Américas Puebla

Abstract

Observational evaluations of mathematics support can be distorted when attendance accumulated across different academic periods is attached to a focal course outcome and when grades awarded in different instructional settings are treated as directly comparable. We address both measurement problems in an administrative study of 6,627 students taking their first eligible attempt at a first-year university mathematics course during periods with observable support-centre coverage. Support registrations are aligned to the same academic period as the focal course, performance is standardised within instructor–course–period classrooms, and visit count is modelled nonlinearly rather than assumed to represent a simple treatment dose. Overall, 18.6% of students recorded at least one same-period CMAT registration. Mean classroom-relative performance was −0.060 standard deviations among students with no visits, compared with 0.218 for 1–2 visits, 0.325 for exactly 3 visits, and 0.342 for 4 or more visits. Games–Howell contrasts separated each positive-use group from non-users, whereas none of the contrasts among the three positive-use groups was statistically distinguishable at the 5% level. Classroom fixed-effect models produced the same qualitative pattern, with estimated contrasts of 0.318, 0.443, and 0.468 standard deviations relative to zero visits. The contribution is therefore not a new claim that mathematics-support users outperform non-users, but a cleaner measurement and interpretation of that association: after aligning exposure temporally and comparing students within a common grading context, the dominant pattern is formal engagement versus non-engagement rather than a clearly graded dose-response among users. Because attendance remains voluntary and important determinants of both help-seeking and achievement are incompletely observed, the estimates are associational rather than causal.

Keywords: mathematics support; academic help-seeking; tutoring; first-year mathematics; academic performance; observational study

1 Introduction

Mathematics learning support has become an established component of higher education provision, particularly where students enter mathematically demanding programmes with heterogeneous prior preparation. Reviews of mathematics and statistics support describe a field that has expanded from remedial provision toward broader, institutionally embedded forms of academic support, while also noting persistent difficulty in evaluating impact because participation is usually voluntary and definitions of attendance vary across studies (Lawson et al. 2020; Mullen et al. 2024). The resulting identification problem is substantive rather than merely statistical, since students who seek help may differ from non-users in prior attainment, motivation, perceived difficulty, study effort, time availability, or willingness to use institutional resources.

A positive association between mathematics-support use and academic performance is not itself novel. Studies at mathematics support centres have repeatedly reported higher grades or pass rates among users while acknowledging the difficulty of separating support from selection into help-seeking (Jacob and Ní Fhloinn 2019; Mac an Bhaird et al. 2009). In particular, Jacob and Ní Fhloinn (2019) showed that students who attended once should be distinguished from students who never engaged, reporting a significant difference between those groups even after considering prior mathematical achievement and module studied. The relevant question for a new observational study is therefore not whether any recorded use can be separated from non-use, but whether the exposure and outcome are measured in a way that makes the resulting comparison more interpretable.

The broader help-seeking literature gives this distinction a useful conceptual basis. Academic help-seeking is generally treated as a self-regulated and motivated strategy rather than a passive treatment received by otherwise exchangeable students, and its relationship with achievement varies with the form of help-seeking and student characteristics (Fong et al. 2023). A count of support-centre registrations should consequently not be read automatically as a production-function measure in which each additional visit constitutes a comparable increment of tutoring. It may instead record a behavioural transition into formal help-seeking, followed by heterogeneous patterns of subsequent use.

This study examines administrative data from the Centro de Aprendizaje de Matemáticas (CMAT) at Universidad de las Américas Puebla and makes three related measurement contributions. First, support use is reconstructed contemporaneously, so that only registrations occurring in the same academic period as a student's first eligible attempt at Matemáticas Universitarias (MU) contribute to the exposure. Second, the outcome is expressed relative to the immediate grading environment by standardising the final course result within an instructor–course–period classroom. Third, visit count is analysed nonlinearly in order to test whether the observed association resembles a graded dose-response or is concentrated in the distinction between no recorded use and positive use. None of these choices eliminates student-level self-selection, but together they remove two avoidable sources of measurement contamination—historical accumulation of visits and cross-classroom grading heterogeneity—while making the functional form of the attendance association explicit.

We therefore ask two questions. How is same-period CMAT use associated with classroom-relative performance in first-attempt MU, and, conditional on observing an association, does performance continue to increase systematically across positive levels of use? The second question is especially important in this institutional setting because three registrations are administratively salient through a participation incentive, which means that the distribution of visit counts is partly shaped by institutional architecture rather than by unconstrained demand for tutoring alone.

2 Institutional context

CMAT provides mathematics advising that students may use during the academic period. Students register their contact with the centre, and a registration is the observable unit in the administrative data. A registration does not record tutoring duration, the substantive content of the interaction, or the amount of independent study that follows it; the same student may also register on multiple occasions and work with different advisers. Visit count should therefore be interpreted as a count of observed formal help-seeking contacts rather than as measured tutoring time.

MU is a first-year university mathematics course, and CMAT use during this course occurs within an institutional incentive structure. Three CMAT registrations have historically been relevant to the first-semester PPA participation requirement. This threshold can affect attendance decisions and makes exactly three visits institutionally meaningful, but it does not generate random assignment. We consequently use the threshold to describe the behavioural environment and to motivate nonlinear exposure categories, not as a regression-discontinuity design or as evidence that three visits constitute a biologically meaningful dose. If future data establish exogenous variation in eligibility for the PPA incentive across programmes, students, or periods, that variation could motivate a separate encouragement-type design; no such identifying assumption is made here.

The registration system is also central to the temporal measurement problem. A student's total number of CMAT contacts over several academic periods can include support sought before or after the course whose grade is being analysed. Assigning that accumulated total to a focal MU record would therefore mix contemporaneous help-seeking with later or earlier behaviour. The present analysis counts only registrations observed in the same academic period as the focal MU attempt.

3 Data and methods

3.1 Analytical population

The primary population contains one observation per student, defined as the first eligible attempt at MU. Administrative non-attempt records such as equivalencies and revalidations are excluded from the attempt definition, whereas adverse academic statuses are retained in the primary outcome construction. The analysis is restricted to academic periods for which CMAT registration coverage is available, because a zero recorded visit is interpretable as observed non-use only when the registration system covers the corresponding period. There are 7,777 eligible first-MU attempts across all observed periods, of which 6,627 occur during periods with CMAT coverage and form the primary analytical cohort.

The primary cohort spans 190 instructor–period classrooms, equivalent to an average of 34.9 students per classroom. The classroom definition is described below and is used both to construct the relative outcome and to absorb shared grading context in the fixed-effect models. Because the inferential comparison is fundamentally within classroom, classroom composition and exposure overlap are important diagnostics rather than incidental descriptive statistics.

3.2 Exposure

For each student, the primary exposure is the number of CMAT registrations recorded during the same academic period as the focal MU attempt. We use the four groups in Equation (1) for the main presentation.

$$V \in \{0,\ 1\text{–}2,\ 3,\ 4+\}. \tag{1}$$

This specification separates non-use from several positive-use intensities without imposing linearity in visit count and preserves the institutionally salient value of three visits.

The 1–2 category pools two low-count positive-use levels. We evaluate whether that pooling masks a substantively important difference using two one-sided equivalence tests with a margin of ±0.20 classroom-standardised outcome units. The margin is a substantive interpretive benchmark, chosen to represent a difference small enough that pooling the two low-use levels would not materially alter the substantive interpretation on a standard-deviation scale; it is not presented as a preregistered or data-independent threshold. The exact one- and two-visit groups have means of approximately 0.222 and 0.209, respectively, with an estimated difference of −0.013 and a 95% interval from −0.135 to 0.108; the TOST rejects differences at or beyond the chosen margin ($p = 0.0013$).

The PPA incentive requires additional caution when interpreting visit count. Students may stop at, target, or pass through three registrations partly because of the participation requirement, so the categories are not naturally ordered doses of an otherwise homogeneous intervention. Comparisons involving three and four-or-more visits are therefore descriptive tests of nonlinearity within the observed incentive architecture, not estimates of marginal returns to successive tutoring sessions.

3.3 Outcome and classroom standardisation

The primary outcome is continuous final performance standardised within classroom, as defined in Equation (2). For student i in classroom c,

$$Z_{ic} = \frac{Y_{ic} - \bar{Y}_c}{s_c}, \tag{2}$$

where a classroom is defined as the same instructor teaching the same course in the same academic period. Because the primary analysis contains only MU, the classroom distinction is effectively instructor by academic period. The transformation places each student on a relative scale within the grading context in which the student's work was evaluated.

Classroom standardisation is a measurement choice rather than a claim that all instructor or period differences are error. Raw course grades can reflect differences in assessment, grading practice, cohort composition, or other classroom-level conditions. Standardisation prevents between-classroom location and scale differences from mechanically generating an association between CMAT use and pooled raw grades, while leaving student-level selection into support use unresolved. The resulting Z score should therefore be interpreted as classroom-relative performance, not as a universal measure of mathematical proficiency.

Non-numeric adverse outcomes are retained in the primary specification using the study's adverse-outcome imputation procedure before classroom standardisation. Because this choice can affect the relative position assigned to withdrawn or otherwise non-numeric records, we also report a uniform adverse-imputation sensitivity and a complete-case specification using numeric final grades only.

3.4 Statistical analysis

We first report group-specific means and 95% confidence intervals for Z. Equality of the four group means is assessed with Welch's one-way ANOVA because the Brown–Forsythe diagnostic provides strong evidence of unequal variances. Pairwise comparisons use Games–Howell intervals and multiplicity-aware p-values, which allow unequal group sizes and variances. These tests establish where the observed distribution differs; they

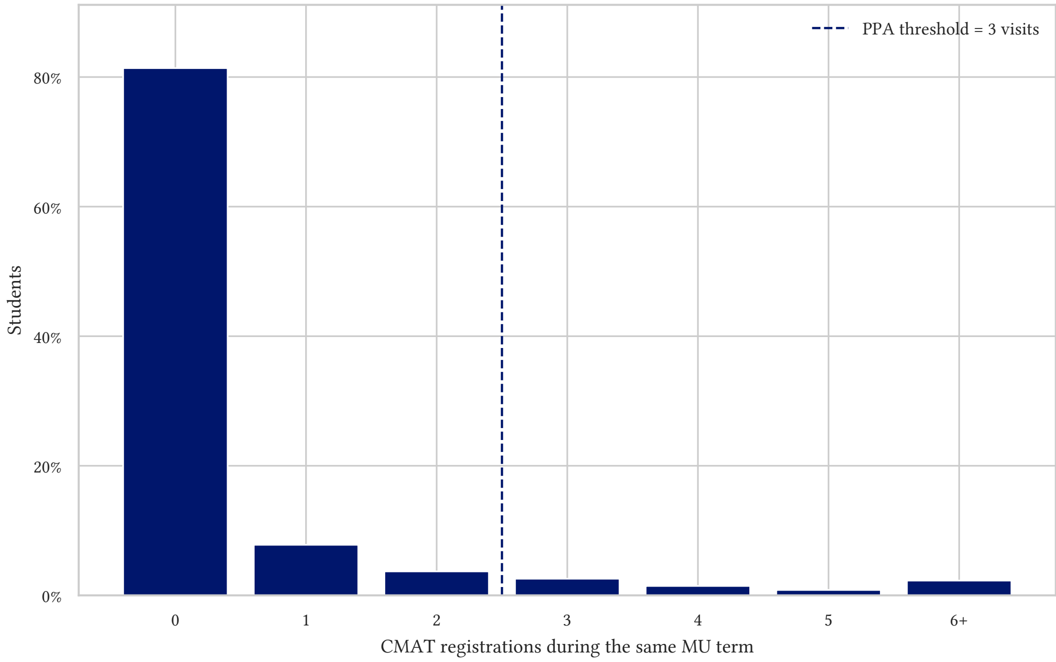

Figure 1: Distribution of same-period CMAT registrations in the first-attempt MU analytical cohort. The dashed line marks the institutionally salient three-visit PPA value for context. The figure is descriptive and does not represent a regression-discontinuity design.

do not, by themselves, determine whether the difference represents support effectiveness, selection into help-seeking, or both.

We then estimate ordinary least-squares models with classroom fixed effects and standard errors clustered by classroom. The principal model includes indicators for 1–2, 3, and 4+ visits, with zero visits as the reference category. The coefficients estimate conditional within-classroom associations and should not be interpreted as causal effects of receiving tutoring. As a secondary upper-tail sensitivity, we compare four or more visits with zero to three visits and repeat that contrast under alternative outcome constructions. The secondary contrast summarizes the robustness of the upper-use association but is not a threshold design.

Propensity-score methods can improve balance on observed pre-exposure characteristics when an appropriate covariate set is available (Austin 2011; Rosenbaum and Rubin 1983). We do not present observed-covariate weighting as a solution to the current identification problem because the available administrative extract lacks sufficiently rich measures of baseline mathematics preparation. The primary interpretation therefore remains associational throughout.

4 Results

4.1 Observed CMAT use in first-attempt MU

Among the 6,627 students in the primary cohort, 5,393 recorded no CMAT registration during the focal MU period, while 1,234 (18.6%) recorded at least one. Exact positive-use counts were 517 students with one visit, 245 with two, 170 with three, and 302 with four or more. Figure 1 shows the resulting distribution, which is strongly concentrated at zero and has a comparatively small positive-use tail. The institutional significance of three visits should be kept in mind when reading this distribution, since bunching near that value can reflect the participation incentive as well as demand for mathematics support.

4.2 Classroom-relative performance by support-use group

Table 1 and Figure 2 show a pronounced separation between students with no recorded CMAT use and each positive-use category. Mean classroom-relative performance was −0.060 for non-users, 0.218 for students with 1–2 visits, 0.325 for exactly three visits, and 0.342 for students with four or more visits.

Table 1: Classroom-relative performance by same-period CMAT-use group.

CMAT visits	N	Mean Z	SD	95% CI
0	5,393	-0.060	1.017	[-0.087, -0.033]
1–2	762	0.218	0.817	[0.160, 0.276]
3	170	0.325	0.724	[0.216, 0.434]
4+	302	0.342	0.712	[0.262, 0.423]

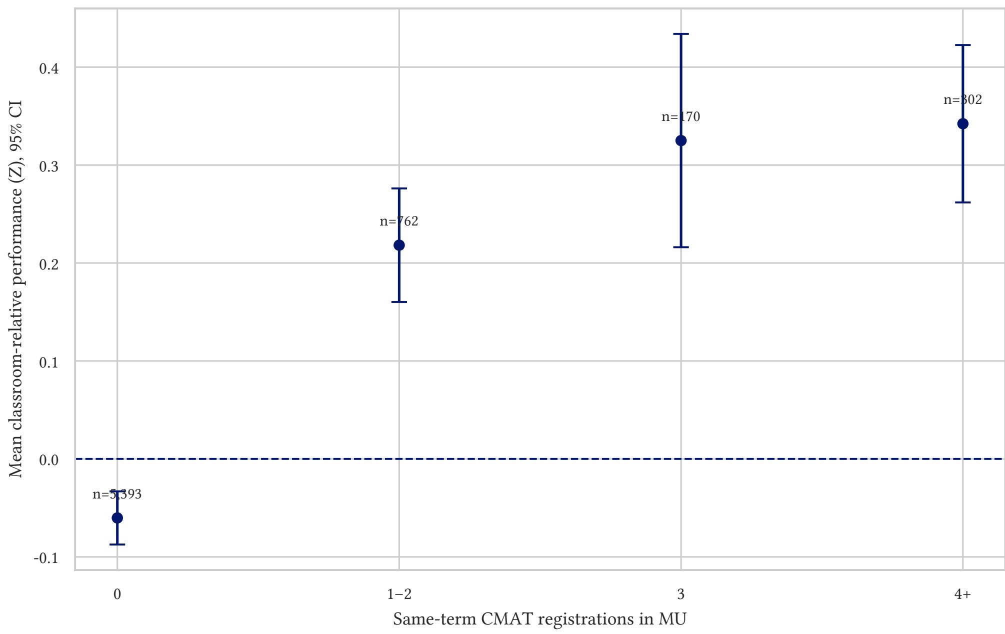

Figure 2: Mean classroom-relative performance by same-period CMAT-use group. Error bars show 95% confidence intervals.

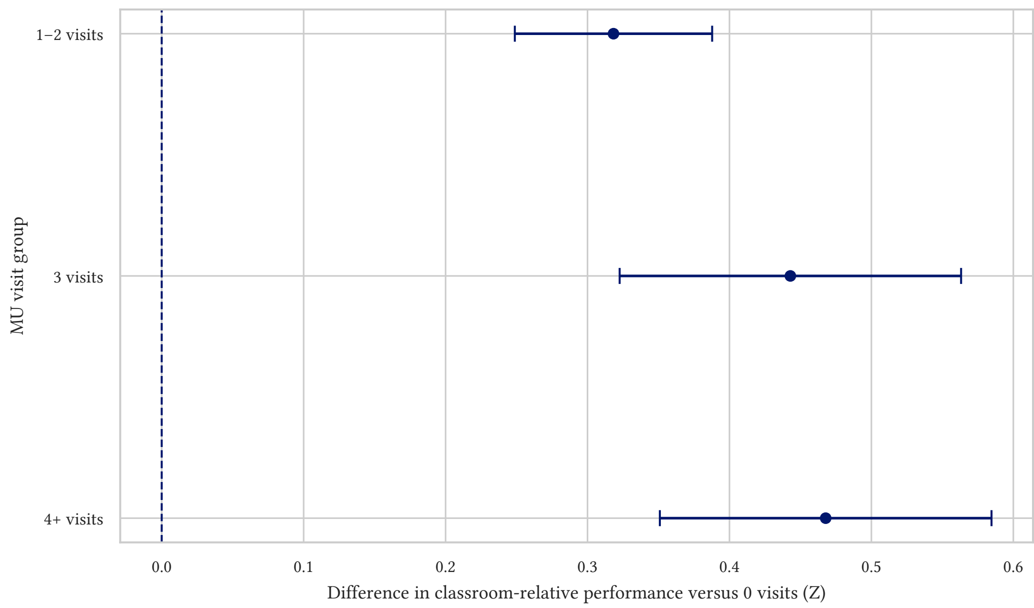

Figure 3: Classroom fixed-effect associations between same-period CMAT-use groups and classroom-relative MU performance, with zero visits as the reference group. Error bars show 95% confidence intervals; standard errors are clustered by classroom.

The Brown–Forsythe test rejected homogeneity of variance ($p < 0.001$), supporting use of the heteroskedastic Welch procedure. Welch's omnibus test also rejected equality of means, $F(3, 546.4) = 56.21$, $p < 0.001$. The pairwise pattern is more informative than the omnibus rejection. Games–Howell comparisons found differences between zero visits and 1–2, three, and four-or-more visits, with adjusted $p \leq 0.001$ in each case. By contrast, comparisons among positive-use groups were not statistically distinguishable at the 5% level: 1–2 versus three visits had $p = 0.327$, 1–2 versus 4+ had $p = 0.069$, and three versus 4+ had $p = 0.900$.

The substantive result is therefore not the omnibus p-value but the shape of the association. The observed gradient is dominated by the separation between no recorded use and positive use, while the data provide much weaker evidence that progressively higher registration counts among users correspond to progressively higher classroom-relative performance. This pattern is consistent with treating recorded CMAT use as an indicator of formal help-seeking behaviour more readily than as a simple tutoring dose.

4.3 Within-classroom associations

Classroom fixed-effect models retain the same qualitative pattern. Relative to students with zero visits in the same classroom environment, the estimated contrasts are 0.318 standard deviations for 1–2 visits, 0.443 for exactly three visits, and 0.468 for four or more visits. All three confidence intervals exclude zero (Figure 3). Because the outcome itself is already standardised within classroom, these models are best understood as an additional restriction of the comparison to students sharing the same instructor-period environment, not as evidence that student-level selection has been removed.

As a secondary upper-tail summary, comparing students with four or more visits against those with zero to three visits yields an unadjusted mean difference of $0.359Z$ (95% CI $[0.277, 0.439]$). A classroom fixed-effect specification estimates $0.388Z$, and adding degree-programme indicators yields $0.375Z$ (95% CI $[0.257, 0.493]$). These estimates show that the upper-use association is not an artifact of the unadjusted comparison, but they do not imply that crossing from three to four registrations causes a discrete increase in performance.

4.4 Outcome sensitivity

The secondary upper-tail contrast is smaller when the analysis is restricted to numeric final grades but remains above zero. The primary adverse-outcome construction gives a mean difference of $0.359Z$ for four-or-more versus zero-to-three visits, the uniform adverse-imputation sensitivity gives $0.343Z$, and the complete-case specification gives $0.211Z$ (95% CI $[0.137, 0.281]$). Figure 4 shows that the magnitude depends on how adverse non-numeric outcomes are handled, reinforcing the decision to treat this binary upper-tail contrast as a sensitivity

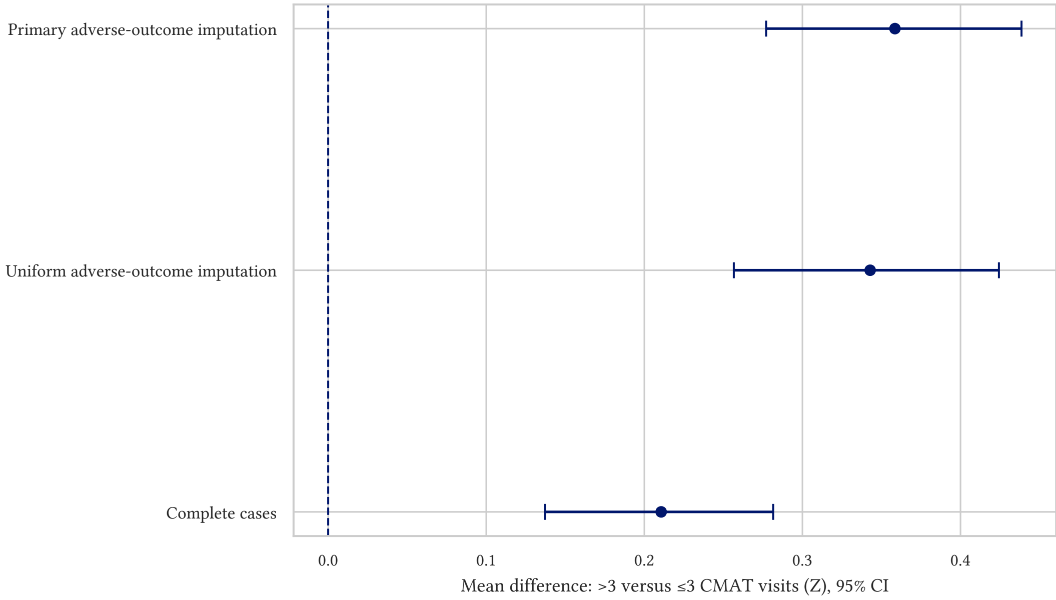

Figure 4: Sensitivity of the secondary four-or-more versus zero-to-three visit association to alternative constructions of the continuous outcome. This comparison is secondary to the four-group analysis.

rather than the main estimand.

5 Discussion

The central contribution of this study concerns measurement and interpretation rather than the rediscovery of a positive user–non-user difference. Previous mathematics-support research has already shown that even one visit can distinguish students who engage with support from students who never attend (Jacob and Ní Fhloinn 2019). Our results instead show that a sizeable positive association remains after two common sources of contamination are removed: support contacts are aligned to the academic period of the focal course, and academic performance is expressed within the grading environment shared by students in the same instructor-period classroom. Once visit count is allowed to be nonlinear, however, the apparent dose pattern is much less compelling than the separation between non-use and positive use.

This interpretation connects naturally to the broader literature on academic help-seeking. Fong et al. (2023) characterise help-seeking as a self-regulated and motivated strategy whose relationship with achievement depends on the form of help-seeking and on student characteristics. In that framework, the present results are more informative about the behavioural state of observed formal help-seeking than about a production function for tutoring sessions. A student who registers once has crossed from non-use into formal engagement, while subsequent registrations may reflect heterogeneous need, persistence, confidence, scheduling constraints, adviser recommendations, or the institutional PPA incentive. Administrative visit count alone cannot distinguish those mechanisms.

The measurement choices also clarify what the study does not solve. Temporal alignment prevents later or earlier support contacts from being misclassified as exposure to the focal MU course, while classroom standardisation prevents pooled grading differences from mechanically generating the result. Neither adjustment makes users and non-users exchangeable. A favourable association can arise because support is beneficial, because students who seek help also engage in other productive behaviours, because students respond to academic difficulty in heterogeneous ways, or through combinations of these mechanisms. Negative selection is also plausible if students experiencing greater difficulty are more likely to seek support, in which case a crude user–non-user comparison could understate an underlying benefit. The available administrative data do not identify the balance among these pathways.

The PPA incentive further weakens a literal dose interpretation. Three registrations are institutionally meaningful, so visit count is partly generated by an incentive architecture rather than solely by a student's preferred amount of tutoring. This feature is analytically useful because it helps explain the shape of attendance, but it is not exogenous in the present design. A future encouragement analysis would require credible evidence that variation in PPA exposure or eligibility is externally determined with respect to potential academic outcomes.

6 Limitations

Several limitations constrain interpretation. First, the study lacks comprehensive pre-exposure measures of mathematics proficiency, and classroom fixed effects do not address student-level self-selection into formal help-seeking. Second, CMAT records count registrations rather than tutoring duration or content, and manual registration can introduce duplicate or missing contacts. Third, visits and course performance are observed within the same academic period, so help-seeking can respond to evolving academic difficulty; temporal alignment improves measurement relevance but does not establish temporal exogeneity. Fourth, treatment of non-numeric adverse outcomes affects estimated magnitudes, although the sensitivity analyses preserve the direction of the association. Fifth, the institutional PPA incentive can shape the distribution of observed visit counts, so comparisons across positive-use categories should not be interpreted as unconstrained marginal responses to additional tutoring. Finally, although the primary models absorb a common grading context, stronger reporting of classroom size and within-classroom exposure overlap is needed to document precisely where the fixed-effect contrasts are identified.

7 Ethics, data governance, and reproducibility

The study uses institutional administrative records, and no row-level student data are reported in the manuscript, figures, or aggregate tables. Raw administrative microdata are not suitable for public release because they are subject to institutional privacy and data-governance restrictions. The analytical code, cohort-construction logic, software-environment metadata, and manuscript build process are maintained under version control, making it possible to release a publication-specific replication package containing code and non-disclosive aggregate or synthetic materials without distributing protected student records.

8 Conclusion

For first-attempt university mathematics, contemporaneous CMAT use is strongly associated with higher classroom-relative academic performance, but the main contribution is not the existence of that association. By aligning support contacts to the focal course period, comparing performance within a common grading environment, and modelling visit count nonlinearly, the analysis shows that the most stable feature of the data is the separation between no recorded formal help-seeking and positive use, whereas differences among positive-use intensities are much less clearly ordered. This pattern does not identify a causal effect of tutoring, yet it provides a more defensible observational description of mathematics-support use by separating measurement improvement from causal identification and by avoiding a simple dose-response interpretation that the institutional data do not support.

References

Austin, P. C. (2011). "An Introduction to Propensity Score Methods for Reducing the Effects of Confounding in Observational Studies". In: *Multivariate Behavioral Research* 46.3, pp. 399–424. DOI: 10.1080/00273171.2011.568786.

Fong, C. J., C. Gonzales, C. Hill-Troglin Cox, and H. B. Shinn (2023). "Academic Help-Seeking and Achievement of Postsecondary Students: A Meta-Analytic Investigation". In: *Journal of Educational Psychology* 115.1, pp. 1–21. DOI: 10.1037/edu0000725.

Jacob, M. and E. Ní Fhloinn (2019). "A quantitative, longitudinal analysis of the impact of mathematics support in an Irish university". In: *Teaching Mathematics and its Applications* 38.4, pp. 216–229. DOI: 10.1093/teamat/hry012.

Lawson, D., M. Grove, and T. Croft (2020). "The evolution of mathematics support: a literature review". In: *International Journal of Mathematical Education in Science and Technology* 51.8, pp. 1224–1254. DOI: 10.1080/0020739X.2019.1662120.

Mac an Bhaird, C., T. Morgan, and A. O'Shea (2009). "The impact of the mathematics support centre on the grades of first year students at the National University of Ireland Maynooth". In: *Teaching Mathematics and its Applications* 28.3, pp. 117–122. DOI: 10.1093/teamat/hrp014.

Mullen, C., E. Howard, and A. Cronin (2024). "A scoping literature review of the impact and evaluation of mathematics and statistics support in higher education". In: *Educational Studies in Mathematics* 117, pp. 1–22. DOI: 10.1007/s10649-024-10332-6.

Rosenbaum, P. R. and D. B. Rubin (1983). "The central role of the propensity score in observational studies for causal effects". In: *Biometrika* 70.1, pp. 41–55. DOI: 10.1093/biomet/70.1.41.