Humanities and Social Sciences Communications

Article in Press

https://doi.org/10.1057/s41599-026-07835-3

Exploring subgroup heterogeneity in online and in-person learning outcomes using causal forest

Received: 7 June 2025

Accepted: 26 May 2026

Cite this article as: Hattab, Z., Aqel, F.A.-Z., Kharousha, A. *et al.* Exploring subgroup heterogeneity in online and in-person learning outcomes using causal forest. *Humanit Soc Sci Commun* (2026). https://doi.org/10.1057/s41599-026-07835-3

Zaid Hattab, Fatima Al-Zahra' Aqel, Abdalmenem Kharousha, Mohammed Yasin & Yahya Jaafra

Exploring Subgroup Heterogeneity in Online and In-Person Learning Outcomes Using Causal Forest

Abstract

This study examines the heterogeneous effects of online versus face-to-face (F2F) learning on student outcomes, using machine learning to uncover subgroup differences rather than focusing on average effects. Analyzing data from 2,210 students, we apply a causal forest method to estimate individualized and subgroup-level effects, considering factors like prior academic performance, specialization, and instructor differences. Nine out of 35 subgroups, defined based on baseline student characteristics (e.g., academic performance, major, and instructor), exhibit significantly heterogeneous effects, with four remaining distinct after multiple testing corrections. For instance, students with lower secondary school scores benefit from online learning in terms of exam scores, while those with higher scores and certain specializations, such as Computer Engineering, experience negative effects. These findings underscore that evaluations of teaching modalities must account for students' heterogeneous characteristics rather than assuming uniform impacts. By identifying which students benefit most from each mode, this research offers realistic insights for optimizing learning outcomes and addressing inequities in diverse educational contexts. These insights challenge one-size-fits-all approaches to digital education and highlight the need for modality selection frameworks that consider the potential heterogeneity. Mainly, our results demonstrate that discipline-specific effects are non-trivial, with quantitative outcomes revealing significant performance declines in applied fields (e.g., Computer Engineering) under online instruction—suggesting that hands-on disciplines may benefit from face-to-face components to preserve learning quality.

1. Introduction

Before 2020, remote education had already developed considerably in many high-income countries, such as USA, where large-scale virtual schools and fully online universities have been operating for years (Betts et al., 2021). However, online learning efforts were limited and timid in many developing or resource-constrained countries (Ndibalema, 2022). While some academic units had begun experimenting with blended learning, many of them were still stuck in the old procedures until the COVID-19 pandemic disrupted the world (Singh et al., 2021). This crisis forced the closure of schools, universities, and other educational institutions. However, virtual classrooms emerged almost immediately, allowing education to continue in an alternative form.

This shift underscored the unprecedented flexibility of online education, extending access to students who were unable to attend their schools and universities, whether due to living remote areas or being affected by conflicts or wars. It appears that distance learning methods have now become a permanent part of our educational system, especially in the fields of engineering and science (Tolonen et al., 2023).

Education in developing regions, such as in Palestine, faces significant challenges stemming from violence, conflict, economic hardship, and checkpoint barriers. The conflict exacerbates these issues by creating an unsafe learning environment through violence, restricting access to schools and universities (Shraim & Khlaif, 2010), and demolishing educational institutions (Smith & Scott, 2023). Additionally, difficult economic conditions and rising poverty rates have led to high school dropout rates, inadequate educational resources, and overcrowded classrooms further hindering student-teacher interaction and increasing reliance on the internet. However, frequent disruptions to internet connectivity in the region due to external constraints create additional barriers to digital learning, exacerbating these challenges (Smith & Scott, 2023; Palestinian Ministry of Higher Education, 2024). These challenging conditions prompted us to conduct an in-depth study of e-learning and its differences from face-to-face education, aiming to identify strengths, weaknesses, and actionable improvements to address urgent needs for effective solutions.

The transition to online education during emergencies has highlighted a significant question on how online education compares to traditional face-to-face (F2F) learning in terms of equality and effectiveness (Abbas et al., 2022). Teachers play a crucial role in addressing these disparities, drawing on their experiences during the COVID-19 pandemic to adapt their teaching methods and support diverse learners (Y. Huang et al., 2023). Additionally, students who experienced the initial shift to online learning are now sharing their insights with newer generations, creating a continuum of adaptation and improvement. However, further research is needed to explore how these dynamics contribute to reducing inequality and enhancing the effectiveness of online education (Alarifi & Song, 2024).

Recent studies provide valuable insights into the comparative effectiveness of different teaching methods. For instance, research comparing F2F and e-learning found that while F2F settings yielded better immediate learning outcomes due to direct interaction and social engagement, e-learning offered flexibility, allowing students to learn at their own pace, which benefited those needing extra time to grasp concepts (Abidin et al., 2023) . However, knowledge retention rates were higher in F2F environments, and students reported greater motivation and interest in F2F classes due to collaborative activities (Nemetz et al., 2017). In contrast, e-learning was praised for its convenience, but often criticized for being less engaging and isolating (Norman, Ph.D., 2020; Rajabalee et al., 2020) .These findings align with prior meta-analytic evidence indicating that online and blended learning environments can achieve outcomes comparable to or exceeding those of traditional face-to-face instruction under certain conditions (Bernard et al., 2014).

Similarly, a study comparing video-based learning (VBL) and F2F instruction reported that VBL was equally effective for skill acquisition and even superior in cognitive assessments, despite limitations like reduced real-time feedback (Eidenberger & Nowotny, 2022) .

These findings underscore the need to move beyond average effects and examine how different student subgroups are impacted by each mode. Recent educational research similarly emphasizes that reliance on aggregate, variable-centered analyses may obscure meaningful subgroup differences and individual learning patterns (Saqr et al., 2024). Longitudinal evidence further suggests that students follow distinct and relatively stable learning trajectories that are strongly linked to engagement and final achievement (Gao et al., 2025), reinforcing the importance of explicitly modeling heterogeneity in educational contexts. However, prior work often lacks sufficiently large sample sizes and integration of quantitative outcomes (e.g., test scores, retention) with contextual covariates (e.g., socioeconomic status, infrastructure).

This paper makes three contributions to the literature on instructional modalities. First, we leverage causal forest machine learning methodology to estimate the heterogeneous effects of online versus F2F learning, moving beyond conventional approaches by focusing on subgroup disparities rather than aggregate outcomes through analysis of an institutional sample (N=2210). Second, by identifying which students benefit most from each mode, we advocate for tailored interventions and hybrid models. Third, we emphasize the critical need for of replicating the comprehensive F2F ecosystem—including essential library resources, counseling services, and social supports—in online environments, guided by Universal Design for Learning (UDL) principles to ensure inclusivity and accessibility (Barbour et al., 2020). In general, this work provides both a template for equitable policy design and a framework for UDL-aligned implementation across diverse educational contexts.

2. Data

The data for this study were obtained in coordination with the office of the vice president for academic affairs at a Palestinian university. Formal authorization was granted to access institutional records through the university's computer center, in strict adherence to ethical standards and privacy protocols. All identifying information was anonymized using coded identifiers to ensure the confidentiality of both students and faculty. The processed and fully anonymized dataset is available through https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PMRYUH **.** The processed student-level dataset is provided under restricted access in accordance with institutional data protection regulations. Raw institutional records cannot be made publicly available.

The initial dataset included records for approximately 3,100 students enrolled in the obligatory Calculus I course. After removing incomplete records and filtering out students who were not in their first year (who accounted for only ten complete cases and were mostly course repeaters), the final analytical sample comprised 2,210 first-year students from a wide range of academic disciplines (1054 in the F2F group versus 1156 online group). The dataset is observational and includes only those who completed all course assessments.

Two instructional modalities are represented: a face-to-face (F2F) cohort from the first semester of the 2022/2023 academic year and an e-learning via "Zoom" cohort from the first semester of 2023/2024. Importantly, the assessment structure remained consistent across both cohorts: 30% of the final grade was allocated to a midterm exam, 20% to assignments, projects, and activities, and 50% to the final exam. While the structure and grading criteria were identical, the mode of assessment delivery differed: students in the online cohort completed online exams, whereas those in the face-to-face cohort took paper-based exams. This distinction reflects an inherent feature of each instructional modality, as assessment delivery forms part of the overall learning process being compared.

The dataset includes students from diverse majors such as Computer Engineering, Information Technology, Physics, Mathematics, and others. Key covariates include academic major, high school score, gender, number of absence days, exam completion status (on-time or incomplete), and anonymized instructor identifiers. These data were extracted from official university records, which covered course registration, performance, and administrative details. Table 1 summarizes the key variables included in the analysis. In total, 35 prespecified subgroups were defined based on these covariates and are presented in Figure 1. All the variables were selected based on theoretical relevance, prior evidence (Figlio et al., 2013; Hart et al., 2018; Xu & Jaggars, 2013), and data availability. The primary outcome was the final course score in Calculus I, measured on a scale from 35 to 100. The scores (mean = 65.6, SD = 17.0, median = 67.0) show a broad and approximately symmetric distribution with no evidence of ceiling compression, as illustrated in Appendix Section 4.

Data analysis was conducted using the R software environment (R Core Team, 2023). The dataset is considered high quality due to its completeness, internal consistency, and diversity of student characteristics. The sample (N=2,210) provides a suitable empirical basis for investigating treatment heterogeneity within this institutional context. Moreover, propensity scores are automatically estimated in the causal forest method, helping to mitigate potential selection bias which is caused by the observational nature of the data by balancing covariates across treatment groups.

Table 1. Descriptive Statistics

	Face-to-face (N=1054)	Online (N=1156)	p-value (t-test)	Standardized Difference
Gender, %				
Male	52.8	56.1	0.11	0.068
Female	47.2	43.9	0.11	0.068
Location, %[1]				
Centre	60.9	63.3	0.24	0.050
North	34.0	32.0	0.33	0.042
South	2.6	2.6	0.96	0.002
External	2.6	2.1	0.45	0.032
Incomplete final exam, %[2]	5.2	13.8	0.001	0.297
Teacher, %				
Teacher 1	3.2	3.9	0.40	0.036
Teacher 2	21.3	15.4	0.001	0.154
Teacher 3	4.1	7.6	0.001	0.151
Teacher 4	10.0	3.0	0.001	0.287
Teacher 5	4.8	4.0	0.33	0.042
Teacher 6	5.5	4.0	0.09	0.072
Teacher 7	14.9	0.0	0.001	0.592
Teacher 8	10.0	16.4	0.001	0.189
Teacher 9	9.0	5.4	0.001	0.142
Teacher 10	2.9	10.0	0.001	0.291
Teacher 11	0.0	16.2	0.001	0.621
Teacher 12	10.2	5.1	0.001	0.194
Teacher 13	3.8	4.0	0.82	0.010
Teacher 14	0.0	5.0	0.001	0.325
Department, %				
Civil or architecture	15.5	14.4	0.50	0.029
Computer Engineering	38.7	36.9	0.37	0.038
Industrial and Chemical Engineering	5.6	8.0	0.02	0.097
Physics	1.9	0.9	0.04	0.089
Information Technology	31.9	35.1	0.11	0.069
Mathematics	2.1	2.2	0.90	0.005
Other	4.4	2.5	0.02	0.102
Failed before, %	7.6	11.6	0.001	0.139
High school score	86.3	86.2	0.95	0.003
Absence (hours)	0.91	0.29	0.001	0.487

[1] Location represents the student's residence place where: Center indicates residing within the same municipality or immediate catchment area as the university, North for students from counties/districts north of the university's geographic area, South for students from counties/districts south of the university's, and External for Students residing outside national borders.
2 Incomplete exam is a binary variable identifies students who did not complete their exam on the originally scheduled date but took it during an approved makeup period..

3. Methodology

Each student i $(i = 1, \ldots, 2210)$ is assigned a binary indicator (D_i) denoting the educational method received, where $D_i = 1$ corresponds to the online education method, and $D_i = 0$ corresponds to the face-to-face education method. Additionally, each student is characterized by a vector of covariates (x_i), which may function as effect modifiers. These covariates are detailed in Table 1. The observed outcome (Y_i) represents the final score in Calculus 1, measured on a scale of 35 to 100.

Three methods were employed to estimate the average effect of online education relative to face-to-face education. First, a simple t-test was conducted to compare the mean outcomes $(\mu_{online} - \mu_{F2F})$ between the two education groups without accounting for potential confounders. However, given the observational nature of the data, it was necessary to incorporate the covariates (x_i). To achieve this, the following linear regression model was fitted:

$$Y_i = \beta_0 + \beta_1 D_i + \sum_{j=1}^{k} \alpha_j x_{ij} \qquad (1)$$

Here, k represents the number of observed confounders, and $\hat{\beta}_1$, the estimated regression coefficient for D_i, captures the average effect of online education. The model parameters were estimated using traditional ordinary least squares (OLS) regression (Hastie et al., 2009).

These traditional methods, however, primarily focus on estimating average effects, potentially masking important variations at the individual student level (Kent et al., 2020). Moreover, these approaches are parametric, relying on specific model assumptions – such as model's linearity, normally distributed errors, and homoscedasticity - that may not always be valid (Hill & Su, 2013). Beyond mean regression, alternative approaches such as quantile regression and quantile treatment effects (Huang et al., 2017; Koenker & Bassett, 1978) have been widely used to study heterogeneity across different points of the outcome distribution. These models are particularly useful when interest lies in how treatment effects vary along the quantiles of the outcome variable (i.e., distributional heterogeneity). However, standard quantile regression models retain linearity in covariates at each quantile and typically require the specification of interaction terms a priori, which can limit their flexibility when the covariates are numerous or when nonlinear relationships are expected (Fakoor et al., 2023). To address these limitations, we utilized the advanced nonparametric machine learning approach known as the causal forest method (Athey et al., 2019).

The causal forest (CF) algorithm builds upon decision trees and random forests (RF), which are traditionally designed for prediction, by adapting them to estimate causal effects. A decision tree (Hastie et al., 2009) partitions the data into subsets (leaves) based on splitting rules that maximize homogeneity in outcomes within each leaf. Specifically, at each node, the tree identifies the covariate and splitting value that minimize variance within the resulting subgroups. This recursive process continues until a stopping criterion (e.g., minimum leaf size) is met, creating a tree structure that groups students with similar outcomes.

While regression trees are flexible and non-parametric, they can be sensitive to specific splits and the choice of covariates, leading to variability across samples (Hastie et al., 2009). Random forests (Breiman, 2001) improve upon this by training multiple trees on random subsets of the data and covariates, aggregating their predictions to increase robustness and reduce overfitting. The causal forest (Athey & Imbens, 2016) extends these concepts to estimate heterogeneous treatment effects, identifying variations in the impact of online versus face-to-face education across different subgroups. Formally, each student has two potential outcomes: $Y_i(0)$, representing their score if they received face-to-face learning, and $Y_i(1)$, representing their score if they had received online learning instead. The primary objective of this study is to estimate τ_i, the difference between these two potential outcomes. However, the core challenge in causal inference lies in the fact that at least one of these outcomes is unobservable (Imbens & Rubin, 2016). Therefore, the parameters to be estimated in each tree in the causal forest are formally defined as follows (Athey & Imbens, 2016):

$$\hat{\tau}_i(x) = \frac{\sum_{i=1}^{n} \alpha_i(x)(Y_i - \hat{\mu}^{(-i)}(X_i))(D_i - \hat{e}^{(-i)}(X_i))}{\sum_{i=1}^{n} \alpha_i(x)(D_i - \hat{e}^{(-i)}(X_i))^2} \qquad (2)$$

where:

1- Out-of-Bag Predictions: the superscript $(-i)$ refers to predictions made using trees that exclude student i during the training (building the data-driven model) process. This ensures unbiased estimates by avoiding overfitting for individual observations.

2- Conditional Mean Outcome ($\hat{\mu}(x)$): $\hat{\mu}(x)$ represents the estimated conditional mean outcome, $\mathbb{E}[Y_i \mid X_i = x]$, and is predicted using a regression forest (a traditional random forest designed for outcome prediction).

3- Conditional Propensity Score ($\hat{e}(x)$): $\hat{e}(x)$ is the estimated conditional propensity score, $\mathbb{P}[D_i = 1 | X_i = x]$¸ which reflects the probability of receiving online learning given the covariates. This is also predicted using a regression forest, and it is essential for observational datasets such as our study.

4- Weights ($\alpha_i(x)$): $\alpha_i(x)$ denotes the weight assigned to each student i, indicating how frequently i appears in the same "leaf" (or group) as the test observation x. These weights determine the contribution of each observation to the parameter estimation.

This formula estimates the individualized treatment effect (ITE) at x by leveraging the difference between observed and predicted outcomes $(Y_i - \hat{\mu}^{(-i)}(X_i))$ and the deviation of treatment assignments from predicted probabilities $(D_i - \hat{e}^{(-i)}(X_i))$. The denominator accounts for the variance in treatment assignment, ensuring the effect is normalized appropriately (Athey et al., 2019; Athey & Wager, 2019; Dandl et al., 2022).

By using out-of-bag predictions for $\hat{\mu}(x)$ and $\hat{e}(x)$, the causal forest mitigates overfitting and improves the robustness of its causal effect estimates. This approach allows the causal forest to effectively estimate heterogeneous treatment effects across different subgroups defined by the covariates (Dandl et al., 2022). Furthermore, the algorithm employs a built-in sample-splitting procedure known as honest estimation, which uses separate subsets of data to determine the tree structure and to estimate the treatment effects within the leaves. This ensures that no single observation is used for both tasks, thereby preventing overfitting and promoting unbiased estimation of treatment effects, even when analyzing the full dataset (Athey & Imbens, 2016).

The treatment effects , $\hat{\tau}_i(x)$, estimated by all individual trees are then aggregated across the forest to produce the final Conditional Average Treatment Effect (CATE) estimate for the covariate profile x In this study, we utilized a forest comprising 20,000 trees to estimate the CATEs. Furthermore, these CATEs were aggregated to compute treatment effects at two additional levels beyond the individual (student) level:

1- Population-Level Effects (ATE): This represents the traditional Average Treatment Effect, calculated as the mean of all individual-level effects across the entire population.

2- Subgroup-Level Effects (GATE): This captures the Group Average Treatment Effect, which is the mean effect for prespecified subgroups defined based on the baseline covariates, as depicted in Figure 1. Both the ATE and GATE were estimated using the Augmented Inverse Propensity Weighting (AIPW) method, a robust doubly robust estimator (Glynn & Quinn, 2010).

It is reasonable to consider that some variables, particularly the teacher identifier, may exhibit nested or hierarchical effects within the data. In the causal forest framework, such higher-level factors are included directly as covariates rather than modeled through explicit multilevel structures. This is standard practice in forest-based estimators of heterogeneous treatment effects (Athey & Imbens, 2016; Dandl et al., 2022; Wager & Athey, 2018) which condition on observed covariates and flexibly capture nonlinear and interaction effects related to teacher or departmental characteristics. To ensure that this specification did not substantially influence the findings, the causal forest model was re-estimated with teacher-level clustering (Appendix Section 3). The results demonstrated that the estimated treatment effects remained stable, suggesting that within-teacher dependence did not significantly affect the conclusions. In the traditional analyses (particularly the linear regression), the teacher identifier was included as a covariate to account for systematic instructor-level differences rather than modeled as a random effect, which is a standard approach when the focus is on treatment effects rather than teacher-level variance (Gelman & Hill, 2007). This framework for the linear regression is also consistent with the robustness of the causal forest results, at least at the average treatment effect (ATE) level. All causal forest analyses have been conducted using grf package in R software (Tibshirani et al., 2022)

4. Results

Despite the observational nature of the data, approximately half of the covariates are balanced between the face-to-face and online groups, as shown in Table 1 (less than 0.05 p-values and less than 0.1 standardized differences that can be seen in the corresponding columns). Furthermore, potential bias from unbalanced covariates has been addressed by estimating the propensity score and incorporating these covariates into the analysis, as detailed in the methodology section. Table 2 presents the estimated overall average effect of online learning compared to the face-to-face method on students' final scores in Calculus I, based on the three methods discussed previously. The results demonstrate consistency across the methods. However, the inclusion of potential confounders appears to reduce the effect of online learning by approximately one unit, as evidenced by the comparison between the T-test results and those obtained from linear regression and causal forest methods.

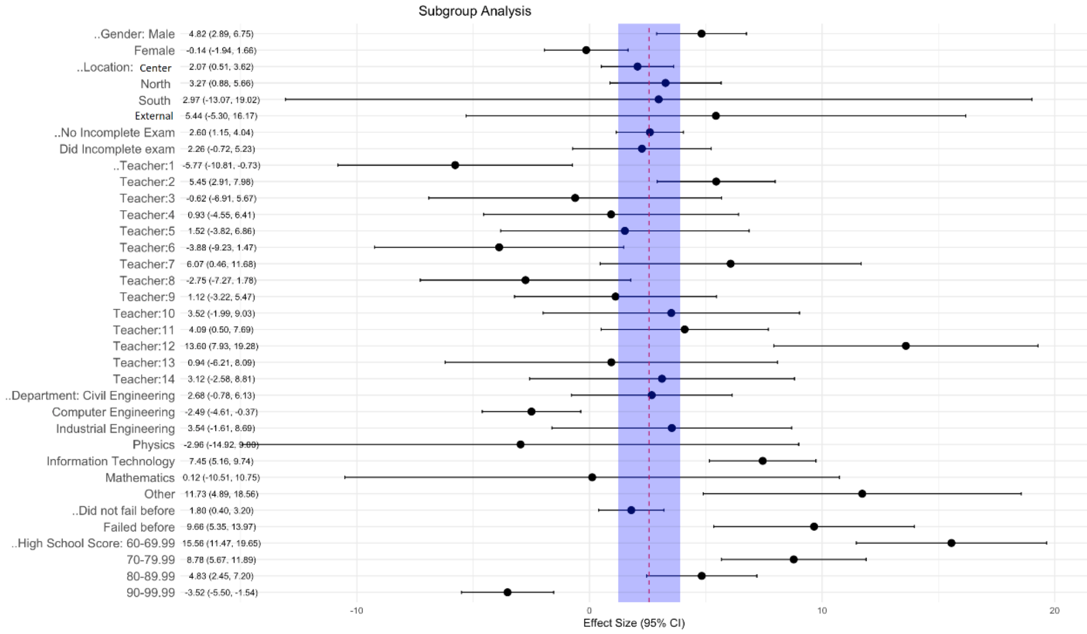
Table 2. Average Effects. (Online Versus Face-to-face)

Method	Simple T-test	Linear Regression	Causal Forest (All covariates)	Causal Forest (Most Important only)[1]
Effect (95% CI)	3.41 (1.98 ; 4.83)	2.67 (1.30 ; 4.04)	2.48 (1.35 ; 3.61)	2.56 (1.23 ; 3.90)
Standardized Effect[2] (95% CI)	0.20 (0.116 ; 0.284)	0.157 (0.076 ; 0.237)	0.146 (0.079 ; 0.212)	0.15 (0.072 ; 0.229)

[1] The most important covariates relate to the covariates that contribute the most in explaining differences in online learning vs. F2F effects; more details are available at Table A1 in the supplementary appendix. [2] Standardized effects were computed by dividing point estimates and confidence interval bounds by the overall standard deviation of the outcome variable (Cohen, 1988) (SD = 17.01).

However, the overall effect appears to mask significant heterogeneity at the subgroup level, as illustrated in Figure 1. Particularly, the impact of online learning is positive for students with lower secondary school scores: [15.6; 95%CI: (11.5, 19.7)] in the 60-to-69.99 subgroup and [8.8; 95%CI: (5.7, 11.9)] in the 70-to-79.99 subgroup. In contrast, the effect is negative for students with higher secondary school scores (90-to-99.99) at [-3.5; 95%CI: (-5.5, -1.5)]. Notably, all these effects are significantly different from the overall effect. (Table A2 in the supplementary appendix provides the tabular representation of these numbers while Figure 1 shows the corresponding forest plot). For interpretability, we also computed the standardized effects (in SD units) for all subgroups, which are presented in Appendix (Section 5) Table A3 and visualized in Figure A5. These standardized estimates facilitate comparison of effect magnitudes across subgroups on a common scale.

Figure 1. Subgroups effects of online versus face-to-face education. The blue area represents the overall population average effect with its 95% confidence interval.



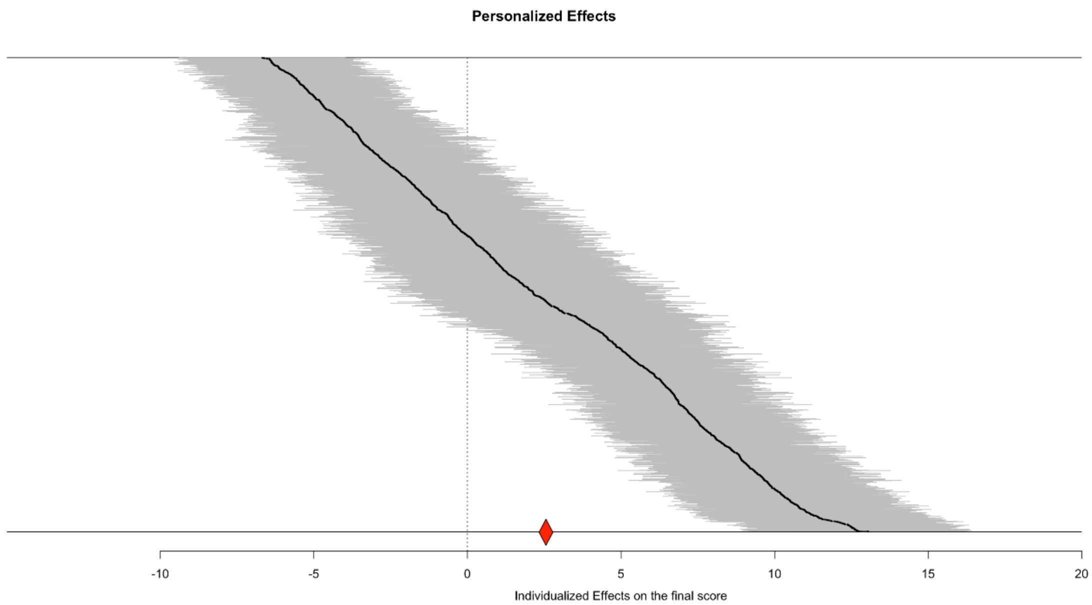
Additionally, specialization appears to act as an effect modifier. For example, students in Computer Engineering are negatively impacted by online learning [-2.5; 95%CI: (-4.6, -0.4)], whereas students in Information Technology benefit positively [7.5; 95%CI: (5.2, 9.8)]. The teacher also emerges as a potential effect modifier, with effects in two teacher-specific subgroups (Teacher 1 and Teacher 12) significantly differing from the overall effect.

Overall, 9 out of 35 subgroups exhibit significantly heterogeneous effects. After applying Bonferroni correction to account for multiple testing, four subgroups remain statistically distinct from the overall average effect: Teacher 12, Computer Engineering, 60-to-69.99 high school score, and 90-to-99.99 high school score subgroups.

This heterogeneity is further evident at the individual student level, as demonstrated by the caterpillar plot in Figure 2. The estimated effects range from **-6.7 to 13**, with many students showing significantly higher or lower effects compared to the overall average, as indicated by the gray confidence intervals.

These findings highlight substantial variability in the impact of online learning, influenced by both subgroup characteristics and individual differences, underscoring the importance of tailoring educational approaches to specific contexts and populations.

Figure 2. Individual-level effects of online versus face-to-face education. The caterpillar plot displays personalized effects (black line), ordered from smallest to largest, with 95% bootstrapped confidence intervals (gray area). These effects are compared to the overall average effect, represented by the red rhombus.



5. Discussion

This study employed causal forest machine learning techniques to investigate heterogeneous treatment effects of online versus face-to-face (F2F) instruction among a sample of 2,210 university students. Moving beyond traditional analyses focused on average treatment effects, we applied a doubly robust estimation strategy using 20,000 decision trees to uncover nuanced variations in Calculus I performance across student subgroups—patterns that may not be captured using traditional regression techniques. The analysis yielded three principal findings: (1) the impact of instructional mode varied significantly by prior academic achievement, with online learning positively affecting lower-performing students (e.g., an average gain of 15.6 points for those in the 60–69.99 percentile) while slightly disadvantaging high-performing students (e.g., a decrease of 3.5 points for those in the 90–99.99 percentile); (2) discipline-specific effects were observed, as Information Technology students benefited from online instruction (+7.5 points), whereas Computer Engineering students exhibited a decline in performance (−2.5 points); and (3) instructor-level variation remained pronounced, illustrated by the disproportionate positive effect associated with a particular instructor (referred to here as Teacher 12). These findings collectively challenge the assumption that a single mode of delivery is equally effective for all learners and underscore the importance of tailoring instructional strategies to individual characteristics.

This study addresses two key gaps in the existing literature on instructional modality comparisons. First, whereas prior research (e.g., (Abidin et al., 2023; Bernard et al., 2014) has largely focused on aggregate treatment effects, our approach highlights how machine learning-based subgroup analyses can reconcile conflicting findings by revealing countervailing effects that are masked when only mean differences are considered. Second, extending the work of (Y. Huang et al., 2023) , who emphasized the importance of teacher adaptability, we quantify the interaction between instructor effects and delivery mode—an issue particularly salient in resource-constrained and conflict-affected contexts such as Palestine. Unlike traditional methodologies that typically treat student and contextual characteristics as confounders (e.g.,Rajabalee et al., 2020), our application of causal forests conceptualizes these variables as effect modifiers. This shift not only facilitates a more nuanced understanding of treatment heterogeneity but also supports the development of targeted, equity-oriented interventions. Our work aligns with the emerging field of precision education (Alarifi & Song, 2024), offering a scalable framework for addressing the dual challenges of access and quality in educational systems under stress.

The study's key findings reveal significant heterogeneity in the effectiveness of online versus face-to-face education, contradicting the belief in a universally superior educational format. While the overall average effect slightly favors online learning, subgroup analyses demonstrate that outcomes depend critically on student characteristics, such as prior academic performance (high school scores), field of study (e.g., Computer Engineering vs. Information Technology), and even instructor differences. Notably, online learning benefits students with lower high school scores but harms high-achievers, suggesting that adaptive educational policies—rather than one-size-fits-all approaches—are needed to maximize learning outcomes. However, caution is warranted when using the final course grade as the sole outcome indicator, as it may not fully reflect students' educational development and could misrepresent online learning's impact—particularly if high-achieving students underperform. This raises the possibility that online modalities may inadvertently hinder advanced learners, raising questions about their broader suitability in some educational settings. These results underscore the importance of personalized education strategies, where institutional decisions about modality should account for student preparedness and disciplinary demands.

The study demonstrates how causal forests reveal hidden patterns in the data missed by conventional methods, enabling targeted interventions—like extra support for struggling online learners or hybrid models for high achievers. Teacher-level differences further highlight the importance of instructional adaptability. These insights can guide policymakers toward more equitable blended learning systems, while researchers can apply similar machine learning techniques to study other educational variables. Moreover, because the causal forest employs an honest estimation procedure with internal sample-splitting, the risk of overfitting is inherently minimized, supporting the stability and internal robustness of the estimated effects. While this study provides valuable insights, several limitations should be acknowledged. First, the observational nature of the data means unmeasured confounders could still influence the results, although the causal forest's built-in propensity score estimation helps mitigate selection bias. In this context, the sample provides a suitable basis for the present analysis, while more fine-grained subgroup estimation may be subject to reduced statistical power and may limit the generalizability of subgroup-specific findings. Second, the online and face-to-face groups were taught in subsequent semesters rather than concurrently. However, the university's grading policies, course structure, and institutional rules remained consistent across semesters, and the same Calculus I curriculum was delivered in both formats, reducing concerns about temporal disparities. In addition, while the assessment policies and grading criteria were consistent, delivery differed by mode (digital vs. in-person). This distinction can be viewed from two perspectives: as a strength, it reflects the comprehensive comparison of complete learning environments, since assessment constitutes an integral component of the educational process; yet, as a limitation, it may have introduced minor mode-related effects that should be considered when interpreting the findings.

The bounded scale of the outcome variable presents a potential limitation. However, the absence of a ceiling effect is supported by the broad and approximately symmetric distribution of scores (range = 35–100; M = 65.6, SD = 17.0), which minimizes concerns that the results were artificially constrained by the measurement scale.

Another limitation is the reduced statistical power in subgroup analyses due to smaller sample sizes when stratifying by covariates (e.g., teacher, major, or high school score). To address this, we applied Bonferroni correction for multiple testing, ensuring robust inference for significant subgroup effects. A further consideration relates to the hierarchical nature of the data, as students were nested within teachers and departments. While the primary models accounted for these factors by including teacher and department identifiers as covariates—a standard approach for estimating average treatment effects—we also examined robustness through a sensitivity analysis incorporating teacher-level clustering (Appendix Section 3). The results confirmed that both the average and subgroup treatment effects remained consistent, with slightly wider confidence intervals, indicating that within-teacher dependence did not substantially affect the conclusions. This stability suggests that the observed heterogeneity reflects systematic differences associated with student characteristics and disciplinary context, rather than arising from the hierarchical data structure. While the present sensitivity analysis indicates that teacher-level clustering did not substantially affect the findings, future research could extend this work by examining clustering effects more comprehensively across different hierarchical levels or educational contexts. In addition, subsequent studies could integrate quantile treatment effect approaches (Huang et al., 2017; Koenker & Bassett, 1978) to explore whether the heterogeneity identified across covariates is also reflected along different points of the outcome distribution, thereby offering a complementary perspective on the robustness and distributional patterns of learning gains.

Finally, while causal forests excel at detecting heterogeneity, their interpretability remains challenging compared to traditional regression models, and the algorithm's performance depends on the quality of covariate selection. Future studies could combine these methods with qualitative research to better understand the mechanisms behind the observed effects.

6. Conclusion

This study contributes to the evidence base on the effectiveness of instructional modalities by applying causal forest machine learning to reveal heterogeneity in treatment effects that traditional average-based analyses may overlook. Three main findings arise: First, the relative effectiveness of online versus face-to-face instruction varies significantly depending on student characteristics. Students with lower prior performance tend to benefit more from online learning, whereas high-performing students and those in specific fields (such as Computer Engineering) may see negative impacts. Second, differences between instructors remain influential across modalities, highlighting the value of flexible teaching approaches—particularly in resource-limited contexts like Palestine, where structural challenges deepen educational disparities. Third, our use of machine learning illustrates how precision education concepts can be translated into implementable policies by identifying which subgroups benefit most from specific modalities.

Although the study relies on observational data, the combination of doubly robust estimation and extensive subgroup analysis offers a strong methodological framework for future research aiming to address both scalability and equity. Nonetheless, limitations such as cohort-based time differences and reduced statistical power in narrowly defined subgroups point to the value of future randomized trials and mixed-method designs. Looking ahead, educational institutions are encouraged to adopt blended models that use online formats to support underserved learners while maintaining face-to-face instruction for more advanced students, accompanied by investments in teacher development and infrastructure aligned with Universal Design for Learning (UDL) principles.

By providing empirical evidence on how teaching modalities differentially affect diverse student populations, this study aligns with the objectives of Sustainable Development Goal 4 (Quality

Education), supporting efforts to promote equitable, inclusive, and context-responsive educational practices in resource-constrained settings.

Overall, the findings move beyond simplistic debates over modality choice and support a more nuanced, data-informed approach that considers the dynamic interaction among students, instructors, and institutional context in shaping educational outcomes.

References

Abbas, M., Pervaiz, M., & Malik, F. (2022). Comparison of Students' Learning through Face-to-Face and Online Classes during Covid-19. *Global Regional Review*, *VII*(II), 190–199. https://doi.org/10.31703/grr.2022(VII-II).18

Abidin, J., Erwina Azizah Hasibuan, & Mahdalia Harahap. (2023). The Effect of Online Learning VS Face-To-Face Learning on Student Learning Motivation Post the Covid—19 Pandemic at Schools and Colleges. *JOURNAL OF DIGITAL LEARNING AND DISTANCE EDUCATION*, *1*(8), 329–337. https://doi.org/10.56778/jdlde.v1i8.76

Alarifi, B. N., & Song, S. (2024). Online vs in-person learning in higher education: Effects on student achievement and recommendations for leadership. *Humanities and Social Sciences Communications*, *11*(1), 86. https://doi.org/10.1057/s41599-023-02590-1

Athey, S., & Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, *113*(27), 7353–7360. https://doi.org/10.1073/pnas.1510489113

Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. *The Annals of Statistics*, *47*(2). https://doi.org/10.1214/18-AOS1709

Athey, S., & Wager, S. (2019). Estimating Treatment Effects with Causal Forests: An Application. *Observational Studies*, *5*(2), 37–51. https://doi.org/10.1353/obs.2019.0001

Barbour, M. K., Labonte, R., Kelly, K., Hodges, C., Moore, S., Lockee, B., Trust, T., & Bond, A. (2020). *Understanding Pandemic Pedagogy: Differences Between Emergency Remote, Remote, and Online Teaching*. https://doi.org/10.13140/RG.2.2.31848.70401

Bernard, R. M., Borokhovski, E., Schmid, R. F., Tamim, R. M., & Abrami, P. C. (2014). A meta-analysis of blended learning and technology use in higher education: From the general to the applied. *Journal of Computing in Higher Education*, *26*(1), 87–122. https://doi.org/10.1007/s12528-013-9077-3

Betts, K., Delaney, B., Galoyan, T., & Lynch, W. (2021). Historical Review of Distance and Online Education from 1700s to 2021 in the United States: Instructional Design and Pivotal Pedagogy in Higher Education. *Journal of Online Learning Research and Practice*, *8*(1). https://doi.org/10.18278/jolrap.8.1.2

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Cohen, J. (2009). *Statistical power analysis for the behavioral sciences* (2. ed., reprint). Psychology Press.

Dandl, S., Hothorn, T., Seibold, H., Sverdrup, E., Wager, S., & Zeileis, A. (2022). *What Makes Forest-Based Heterogeneous Treatment Effect Estimators Work?* (arXiv:2206.10323). arXiv. http://arxiv.org/abs/2206.10323

Eidenberger, M., & Nowotny, S. (2022). Video-Based Learning Compared to Face-to-Face Learning in Psychomotor Skills Physiotherapy Education. *Creative Education*, *13*(01), 149–166. https://doi.org/10.4236/ce.2022.131011

Fakoor, R., Kim, T., Mueller, J., Smola, A. J., & Tibshirani, R. J. (2023). *Flexible Model Aggregation for Quantile Regression* (arXiv:2103.00083). arXiv. https://doi.org/10.48550/arXiv.2103.00083

Figlio, D., Rush, M., & Yin, L. (2013). Is It Live or Is It Internet? Experimental Estimates of the Effects of Online Instruction on Student Learning. *Journal of Labor Economics*, *31*(4), 763–784. https://doi.org/10.1086/669930

Gao, Z., Yan, H., Liu, J., Cui, C., Wang, J., Zhang, X., Sun, X., & Feng, J. (2025). Untangling complexity of competency evolution: Multi-channel learning trajectories of score, engagement and problem-solving efficiency. *British Journal of Educational Technology*, bjet.70017. https://doi.org/10.1111/bjet.70017

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press. https://doi.org/10.1017/CBO9780511790942

Glynn, A. N., & Quinn, K. M. (2010). An Introduction to the Augmented Inverse Propensity Weighted Estimator. *Political Analysis*, *18*(1), 36–56. https://doi.org/10.1093/pan/mpp036
Hart, C. M. D., Friedmann, E., & Hill, M. (2018). Online Course-taking and Student Outcomes in California Community Colleges. *Education Finance and Policy*, *13*(1), 42–71. https://doi.org/10.1162/edfp_a_00218
Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer New York. https://doi.org/10.1007/978-0-387-84858-7
Hill, J., & Su, Y.-S. (2013). Assessing lack of common support in causal inference using Bayesian nonparametrics: Implications for evaluating the effect of breastfeeding on children's cognitive outcomes. *The Annals of Applied Statistics*, *7*(3). https://doi.org/10.1214/13-AOAS630
Huang, Q., Zhang, H., Chen, J., & He, M. (2017). Quantile Regression Models and Their Applications: A Review. *Journal of Biometrics & Biostatistics*, *08*(03). https://doi.org/10.4172/2155-6180.1000354
Huang, Y., Lazarides, R., & Richter, D. (2023). Teachers' adaptations to COVID-19: Perceived preparedness for distance education, frequency of teacher-student contact, and resources in ICT. *European Journal of Teacher Education*, 1–26. https://doi.org/10.1080/02619768.2023.2288554
Imbens, G. W., & Rubin, D. B. (2016). *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction: Book Reviews* (Vol. 84). http://doi.wiley.com/10.1111/insr.12170
Kent, D. M., Paulus, J. K., Van Klaveren, D., D'Agostino, R., Goodman, S., Hayward, R., Ioannidis, J. P. A., Patrick-Lake, B., Morton, S., Pencina, M., Raman, G., Ross, J. S., Selker, H. P., Varadhan, R., Vickers, A., Wong, J. B., & Steyerberg, E. W. (2020). The Predictive Approaches to Treatment effect Heterogeneity (PATH) Statement. *Annals of Internal Medicine*, *172*(1), 35. https://doi.org/10.7326/M18-3667
Koenker, R., & Bassett, G. (1978). Regression Quantiles. *Econometrica*, *46*(1), 33. https://doi.org/10.2307/1913643
Ndibalema, P. (2022). Constraints of transition to online distance learning in Higher Education Institutions during COVID-19 in developing countries: A systematic review. *E-Learning and Digital Media*, *19*(6), 595–618. https://doi.org/10.1177/20427530221107510
Nemetz, P. L., Eager, W. M., & Limpaphayom, W. (2017). Comparative effectiveness and student choice for online and face-to-face classwork. *Journal of Education for Business*, *92*(5), 210–219. https://doi.org/10.1080/08832323.2017.1331990
Norman, Ph.D., Z. (2020). Understanding the Effect of Distance Learning vs. Face-to-Face Learning Experiences on Students' Engagement in Higher Education. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.3734764
R Core Team. (2023). *R: A Language and Environment for Statistical Computing* [Computer software]. R Foundation for Statistical Computing. https://www.R-project.org/
Rajabalee, B. Y., Santally, M. I., & Rennie, F. (2020). A study of the relationship between students' engagement and their academic performances in an eLearning environment. *E-Learning and Digital Media*, *17*(1), 1–20. https://doi.org/10.1177/2042753019882567
Saqr, M., Vogelsmeier, L. V. D. E., & López-Pernas, S. (2024). Capturing where the learning process takes place: A person-specific and person-centered primer. *Learning and Individual Differences*, *113*, 102492. https://doi.org/10.1016/j.lindif.2024.102492
Shraim, K., & Khlaif, Z. (2010). An e-learning approach to secondary education in Palestine: Opportunities and challenges. *Information Technology for Development*, *16*(3), 159–173. https://doi.org/10.1080/02681102.2010.501782
Singh, J., Steele, K., & Singh, L. (2021). Combining the Best of Online and Face-to-Face Learning: Hybrid and Blended Learning Approach for COVID-19, Post Vaccine, & Post-Pandemic World. *Journal of Educational Technology Systems*, *50*(2), 140–171. https://doi.org/10.1177/00472395211047865
Smith, M., & Scott, H. (2023). Distance Education under Oppression: The Case of Palestinian Higher Education. *Education Sciences*, *13*(7), 729. https://doi.org/10.3390/educsci13070729

Tibshirani, J., Wager, S., Athey, S., & Sverdrup, E. (2022). *grf: Generalized Random Forests* (Version 2.2.1) [R package]. https://CRAN.R-project.org/package=grf}

Tolonen, M., Arvonen, M., Renko, M., Paakkonen, H., Jäntti, H., & Piippo-Savolainen, E. (2023). Comparison of remote learning methods to on-site teaching -randomized, controlled trial. *BMC Medical Education*, *23*(1), 778. https://doi.org/10.1186/s12909-023-04759-3

Wager, S., & Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association*, *113*(523), 1228–1242. https://doi.org/10.1080/01621459.2017.1319839

Xu, D., & Jaggars, S. S. (2013). The impact of online learning on students' course outcomes: Evidence from a large community and technical college system. *Economics of Education Review*, *37*, 46–57. https://doi.org/10.1016/j.econedurev.2013.08.001

Declarations

Competing interests

The authors declare no competing interests.

Funding

This study received no external funding.

Data Availability

Replication materials supporting the findings of this study are available in the Dataverse (https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PMRYUH) The processed student-level dataset is provided under restricted access in accordance with institutional data protection regulations. Raw institutional records cannot be made publicly available.

Ethical Approval

This study was reviewed by the Institutional Review Board (IRB) at An-Najah National University and was granted Exempt (Waiver) status (Ref. No. Sci. Feb. 2026/63; 22 February 2026). The IRB determined that the study met the criteria for exemption from full committee review under institutional guidelines governing research involving secondary, fully anonymized administrative data. All procedures were conducted in accordance with applicable institutional regulations and ethical standards.

Informed Consent

The study utilized secondary administrative academic records that were fully anonymized prior to analysis. In accordance with the IRB exemption determination, the requirement for individual informed consent was waived. No personally identifiable information was accessible to the authors at any stage of the study.