# MEASURING TREATMENT EFFECTS OF ONLINE VIDEOS ON ACADEMIC PERFORMANCE USING DIFFERENCE-IN-DIFFERENCES ESTIMATIONS

B. I. Smith[1]*, J. H. Bührmann[2]* & C. Chimedza[3]

## ARTICLE INFO

*Article details*
Presented at the 32nd annual conference of the Southern African Institute for Industrial Engineering (SAIIE), held from 4-6 October 2021 in Muldersdrift, South Africa.

Available online 29 Nov 2021

*Contact details*
* Corresponding author
bevan.smith@wits.ac.za

*Author affiliations*
1 Academic Development Unit, University of the Witwatersrand, South Africa

2 School of Mechanical, Industrial and Aeronautical Engineering, University of Witwatersrand, South Africa

3 School of Statistics and Actuarial Science, University of Witwatersrand, South Africa

*ORCiD® identifiers*
B. I. Smith
https://orcid.org/0000-0002-4881-6661

J. H. Bührmann
https://orcid.org/0000-0003-0657-9933

C. Chimedza
https://orcid.org/0000-0001-8980-1610

*DOI*
http://dx.doi.org/10.7166/32-3-2613

## ABSTRACT

Supplementing student learning with online videos has produced mixed results in respect of improving academic performance. This study proposed that these mixed results, before COVID-19, are due to most of the literature on online videos being observational studies and not taking confounding factors into account. This study applied the difference-in-differences (DID) technique, which measures treatment effects in observational studies. Using student grades data from an engineering mechanics course, the treatment effects of videos were measured 1) on the entire group that used the videos, 2) on those initially with failing and then with passing grades, and 3) on those grouped by grade percentage ranges. It was found that the videos had no effect on the entire group but did significantly affect those who had initial failing grades — specifically, grades of 40% to 49%. A main finding was that initial grades are an indication of how effective the online videos are in improving grades.

## OPSOMMING

Die aanvulling van studente se leerervaring met voorafopgeneemde video's het gemengde resultate in terme van akademiese vertoning gehad. Hierdie studie stel voor dat die gemengde resultate, voor COVID-19, hoofsaaklik te wyte is aan die aard van die literatuur – die meeste studies hieroor was waarnemende studies en het nie verwarrende faktore in ag geneem nie. Hierdie studie pas die verskil-in-verskille tegniek toe; dit meet die verskille in behandeling tydens waarnemende studies. Deur bestaande data van studente se punte van 'n ingenieursmeganika vak te gebruik, is die behandelingseffekte gemeet 1) op die volledige groep wat van die video's gebruik het, 2) op die wat aanvangklik druipkategorie kandidate was wat oorgegaan het na die slaagkategorie, en 3) op groeperings aan die hand van puntreekse. Daar is gevind dat die video's geen invloed op die algehele groep gehad het nie, maar dat dit die wat aanvanklik druipkategorie kandidate was noemenswaardig beïnvloed het, veral vir studente met punte tussen 40% en 49%. Die belangrikste bevinding is dat die aanvanklike punte 'n aanduiding gee van die doeltreffendheid van video's om punte te verbeter.

## 1 INTRODUCTION

The use of supplementary online videos to improve student learning promises much, but in the past, before COVID-19, has produced mixed results. For example, the literature shows videos having a positive impact on performance [1, 2, 3], a negative impact [4, 5], and no impact [6, 7]. The mixed results have been attributed to different learning strategies [5], passive learning [7], attending fewer face-to-face lectures [1, 5, 4], and videos potentially not being effective across all fields of study [3].

Although these may be factors, this study proposes that a reason for the mixed results is that most of the studies are observational [8], not randomised control trials, and therefore were not measured using the appropriate methods. Unlike randomised control trials, participants in observational studies are not randomly assigned to the control and treatment groups. In these studies, students self-selected to participate in an intervention (such as watching online videos). This potentially results in a selection bias [9, 10] and thus in biased treatment effects.

To measure the treatment effect of an observational study such as an academic intervention (in our case, watching online videos), it is crucial that appropriate measurement techniques be used. For observational studies, if the treatment group (the group that watched videos) had overall better (or worse) results than the control group (the group that elected not to watch videos), it cannot be concluded that the intervention caused a better (or a worse) performance, owing to the self-selection. It may be caused by confounding features, such as the treatment group students being more diligent and harder working or more aware of their failing grades. For this reason, the grades of groups that use online videos cannot simply be compared with groups that do not use online videos.

An example is Traphagan *et al.* who, in their observational study, first found that those watching videos performed worse than the no-videos group. Thereafter they controlled for grade point average (GPA) and found that the two groups performed similarly [4].

Based on these findings, and because the authors had only grade information for the case study that follows, it was proposed to use the difference-in-differences (DID) method to measure the treatment effect on student performance of the videos as an intervention. To the authors' knowledge, this method has thus far not been used in the literature to measure the effectiveness of academic interventions. It is proposed because it does not need a large number of features that describe the students: it simply requires the grades of the students before and after the intervention.

## 2 LITERATURE REVIEW

According to Muller *et al.* the belief that multimedia will improve learning has been around for almost a hundred years [11]. Even before COVID-19, the area of multimedia that saw the most growth was reported to be online education videos [12]. Although the use of videos promises much, the literature reports that their impact on student learning varied. A major reason for videos not delivering what they promise is attributed by the literature [5, 7, 13] to superficial or passive learning.

The attitude [14], motivation [15, 16], and attention [17] of the student have also been shown to be important in learning from multimedia. When investigating why videos are or are not effective, the above-mentioned reasons are certainly valid ones.

### 2.1 Observational studies

In general, there are two types of study to measure intervention or treatment effectiveness. Observational studies are those in which the participants (i.e., students) self-select to use the intervention (in this case, videos) [8]. This contrasts with randomised controlled trials, in which participants are randomly assigned to either a control or a treatment group. Although randomised control trials are seen as the 'gold standard' for evaluating the efficacy of an intervention, ethical issues may arise from randomly assigning students to watch or not watch videos [8].

Furthermore, the benefit of observational studies is that the natural behaviour of participants can be observed — i.e., participants will act naturally when using the intervention, and not be affected by being monitored in a study setting. Therefore, simply allowing the students to self-select to watch the videos is seen as the appropriate way to provide this intervention.

Self-selection, however, generally produces biased results when measuring the effectiveness of the treatment (referred to as 'treatment effects' hereafter), owing to confounding factors.

### 2.2 Previous case studies on online videos performance

When using observational studies to study the effects of online videos on performance, Traphagan *et al.* used a one-way multivariate analysis of covariance [4]. Wieling *et al.* and Williams *et al.* performed multivariate regression to measure the effect of online viewing on performance [2,3]. Leadbeater *et al.* used simple correlation to study how the amount of video watching impacts performance. The problem with regression and correlation studies is that they essentially measure correlation and not causal (treatment) effects [7].

When measuring treatment effects in observational studies, the aim should be to balance out the features so that the treatment group and the control group have balanced features. The aim is to create a counterfactual group to compare with the treatment group. For observational studies, the main way of carrying this out is to use propensity score matching [18]; this is beneficial if the study has a large range of features describing the participants and the effects are measured using cross-sectional data. In the absence of a range of descriptive features, the difference-in-differences (DID) method can be used, which uses data captured over time.

### 2.3 Difference-in-differences

The major benefit of DID is that it can measure treatment effects without needing a large range of features. It simply requires data captured over time. DID takes into account changes in outcome (e.g., grades) in order to measure effects. The intuition behind DID can be explained as follows: Consider measuring the change in grades (positive or negative) of the students that self-selected to watch videos (the treatment group). After watching the videos, there could be the erroneous thought that the change in grades was due to the videos (the intervention). However, to measure the videos' effect properly, there needs to be a counterfactual group to compare against — i.e., a similar group that did not receive the intervention (the control group).

The counterfactual control group's average change in grades should be measured and compared with the treatment group's change in grades. Whenever treatment effects are measured, there is the requirement to compare a treatment group with a counterfactual group. The control group is the counterfactual group that aims to explain what the treatment group would have obtained had they not received the treatment.

DID assumes that only the difference in the change in grades between the treatment group and the control group needs to be compared. This is based on the parallel trends assumption [19, 20], which states that the treatment (videos) group, in the absence of the treatment, would have had the same change in grades as the control group (no-videos) over the time period being considered.

The DID method aims to remove or control for the effects of selection bias. In the literature there are examples that use this technique to measure important concerns such as the effect that raising the minimum wage has on employment [21], estimating the impact of training programmes on earnings [22], and how disability benefits affect time off work after injury [23].

Equation (1) provides an intuition for the treatment effect, or DID [24], where D is the DID value, $\overline{\mathrm{y_{tf}}}$ and $\overline{\mathrm{y_{t\imath}}}$ refer respectively to the average grade at the end and at the beginning of the time span for the treatment group. Similarly, $\overline{\mathrm{y_{cf}}}$ and $\overline{\mathrm{y_{c\imath}}}$ refer respectively to the average grade at the end and at the beginning of the time span for the control group. The equation shows that the change in the control group's average grade from the beginning to the end is subtracted from the change in the treated group's average grade over the same time.

$$D \;=\; \left(\overline{y_{tf}} - \overline{y_{t\imath}}\right) - \left(\overline{y_{cf}} - \overline{y_{c\imath}}\right) \tag{1}$$

Although Equation (1) is a simple way to calculate the DID value, it is also required to calculate the p-values to determine whether these values are statistically significant. For this, ordinary least squares (OLS) regression is used, based on Equation (2) [23, 24, 25]; where y refers to the grade (either mid-year or year-end); T is a binary variable indicating whether the grade is at the start or the end of the time period (mid-year, T = 0; year-end, T = 1); S is a binary variable indicating whether the student received treatment (no-videos, S = 0; videos, S = 1). Finally, T * S is an interaction variable. In terms of the coefficients, $\beta_0$ refers to the control group's initial grade; $\beta_1$ indicates how much the control group changes grades over the time span of the intervention; $\beta_2$ is the difference between the initial grade of the treatment group and control group; and $\beta_3$ is the DID coefficient value, which can be used to estimate the prediction of the treatment effect.

$$y \;=\; \beta_0 + \beta_1 T + \beta_2 S + \beta_3 (T * S) \tag{2}$$

The p-value represents the likelihood of the statistical significance of evidence between the predictor and the response [26]. In order to measure whether the OLS regression model is providing a $\beta_3$ coefficient that accurately predicts the treatment effect on student performance, the p-value for the two-sided t-test is calculated. In this case, the null hypothesis is that the $\beta_3$ coefficient is zero [25, 26], which means that the

$\beta_3$ coefficient is not significant — i.e., that there is no treatment effect. If the p-value is below an α=5% significance level, the null hypothesis can be rejected, and the $\beta_3$ coefficient has explanatory power — i.e., it is significant. If the p-value is bigger than α = 5%, the null hypothesis is accepted that the true value of the beta coefficient is actually zero.

## 3 RESEARCH METHODOLOGY

In this study, grades from a first-year engineering mechanics course in 2017 and in 2018 were used to analyse the actual video watch time as a predictor of student performance. The course had an estimated 1 000 students registered per year. Supplementary videos, each typically between five and ten minutes long, were recorded to cover both conceptual and practice content.

The analysis focused on comparing the students' video watch time for the second semester of the course. The reason for focusing on the second semester grades was that no videos were available in the first semester of 2017, and students already had a starting grade to use as reference with which to compare the changes in grades for the difference-in-differences statistics. In our case, the mid-year grades (after the July exam), calculated based on the Semester 1 assessments, were used as the starting grades for the DID comparisons. In 2017, a total of 80 videos were made available for Semester 2, and in 2018, 129 videos. The video hosting platform, Panotpo$^{TM}$ [27], was configured to record the exact number of minutes that each student watched each video.

Because the students self-selected to watch the videos, the study is seen as observational [8]. Grades vs actual video watch time were first directly compared to evaluate whether the results would be similar to those reported in the literature — i.e., that there was no relationship between videos and performance [4, 5, 6, 7] when ignoring confounding features from the self-selection nature of the study. Next, the DID values for the changes in grades over time were compared to measure more appropriately the treatment effect of watching videos.

### 3.1 Video watch time as a predictor of grades

For the direct comparisons, mid-year grades (from the first semester of the course, running February to July) were compared with three major assessments in the second semester: the results of two major tests (September test — covering two textbook chapters; October test — covering one chapter) and the final November exam (covering four chapters). These were also compared with the overall year-end grades for the course, based on a combination of the mid-year results as well as all of the 2nd semester assessments and tutorial tests.

To determine whether there was any association of minutes watched with performance, the following analyses were carried out for each of the three assessments in each year:

1. Plots were generated to show grades versus minutes watched for all of the students for each test or exam.
2. The beta value of an ordinary least squares (OLS) linear regression model between the total number of minutes that videos were watched, and the students' grades were calculated.
3. T-tests were used to calculate the p-value. The p-value was used to determine whether the linear regression model in the previous step could be used as a significant predictor of student performance when comparing the videos and the no-videos groups.

### 3.2 Difference-in-differences (DID)

For the DID comparisons, the mid-year grades were compared with the final November exam. The DID was first computed for the entire cohort of students. The aim here was to see whether the videos had a treatment effect on the entire group that watched them.

Next, DID estimates were also computed, based on whether students initially passed or failed at mid-year. The aim was to see whether the initial performance was an indicator of the treatment effects. Thereafter, the initial mid-year grades were further broken down into ranges of grades from 30% to above 80% in ranges of 10%. The aim here was to see whether different ranges had different treatment effects.

For all of the DID estimates, the p-values were calculated to determine whether the DID was a significant predictor of the treatment effect.

# 4 RESULTS

## 4.1 Analysis of video watching time

Before comparing the performance of the student grades, the amount of video watching time before assessments was analysed. The distribution of the September 2017 test grades compared with the frequency of videos watched is shown in Figure 1. 44 videos covered the topics for this test. The graph reflects an approximately normal distribution with mean = 51.96%, std = 17.00%.

Figure 2 illustrates the independent variable 'minutes watched', which is skewed to the right. Given that skewness, Spearman's correlation was calculated for all of the assessments in 2017, but it consistently confirmed that there was no relationship between the minutes watched and the grades in any of the assessments.

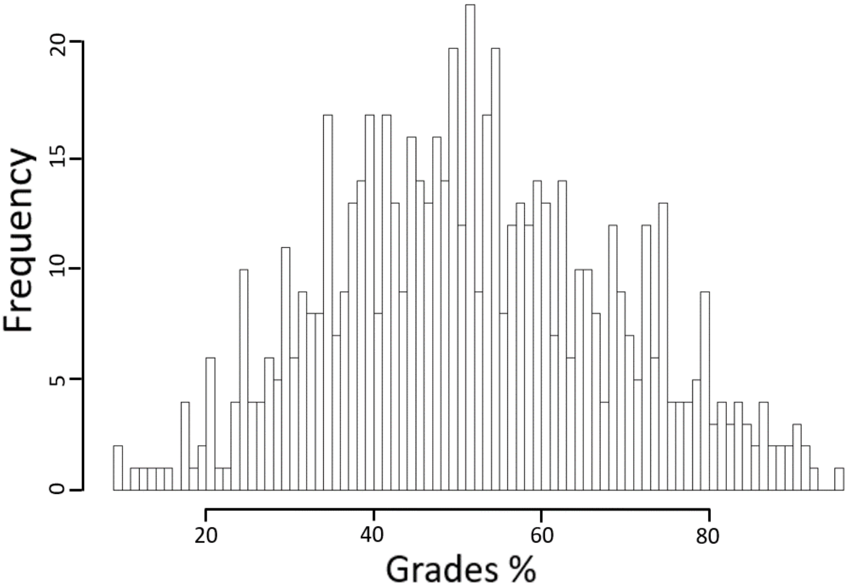


**Figure 1: Histogram of September 2017 test per video frequency**

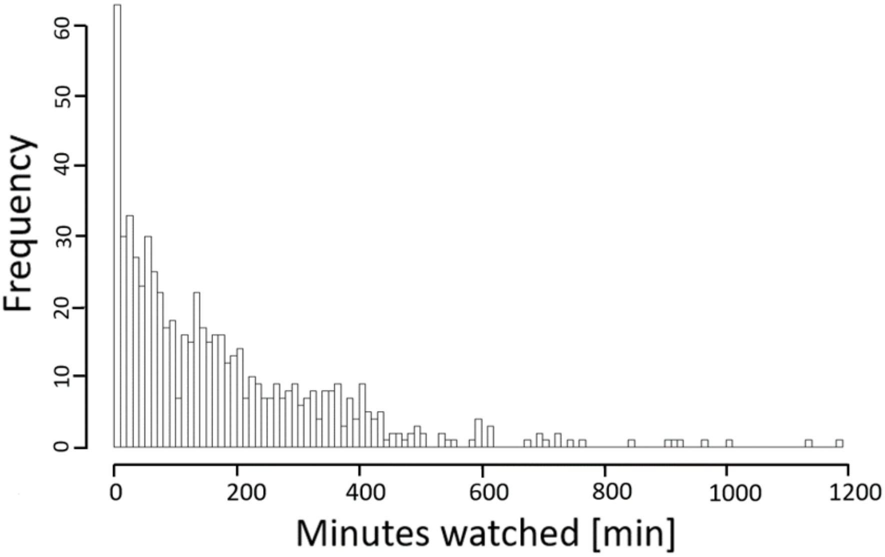


**Figure 2: Minutes watched of September 2017 test per video frequency**

### 4.2 Direct comparison of grades

Figure 3 plots each student's grades for the November 2017 exam against the accompanying minutes of videos watched. There appears to be a random scattering of the data points, indicating no correlation between the number of minutes watched and the grades; in other words, the video watch time is not a predictor of the grades. Scatter plots for the other assessments in 2017 and 2018 gave similar results.

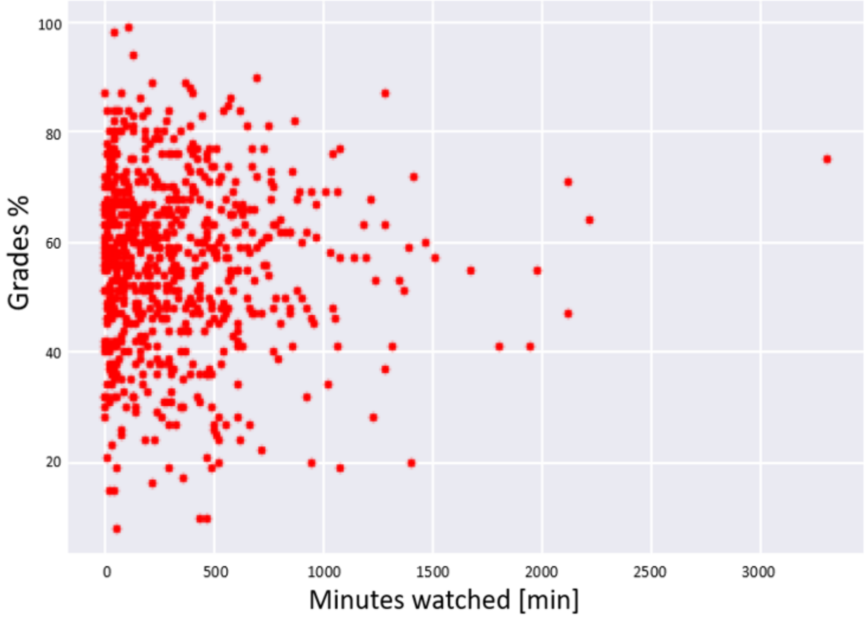


**Figure 3: November 2017 exam — grades of videos group**

To confirm that there is no direct relationship between the number of minutes watched and the grades, ordinary least squares (OLS) regression models were used. Fitting a linear regression model, a beta coefficient and the associated p-value for each assessment were obtained for the students who watched videos. The results are shown in Table 1.

**Table 1: Regression coefficients for minutes videos watched vs assessment grades**

| Assessment | Beta coefficient | p-value | No. of students (videos group) |
|---|---|---|---|
| Sep 2017 test | 0.004 | 0.33 | 652 |
| Oct 2017 test | 0.000 | 0.99 | 499 |
| Nov 2017 exam | - 0.001 | 0.46 | 609 |
| Mar 2018 test | - 0.007 | 0.46 | 236 |
| May 2018 test | - 0.001 | 0.82 | 244 |
| June 2018 exam | 0.486 | 0.52 | 521 |
| Sep 2018 test | 0.003 | 0.37 | 416 |
| Oct 2018 test | - 0.003 | 0.46 | 373 |
| Nov 2018 exam | 0.003 | 0.03 | 529 |

The beta coefficients for the majority of the assessments are close to zero. This is consistent with the null hypothesis, that the beta coefficient is zero and that there is therefore no linear relationship between the minutes watched and the grades. The June 2018 exam is the only assessment that had a beta coefficient significantly higher than 0. This assessment was thus the only one in which a possible linear relationship was found. However, the p-value for this assessment was 0.52, indicating that the linear relationship was not significant at α = 0.05.

The p-values for most of the assessments were also above an α = 0.05 significance level, confirming that there was no relationship between the minutes watched and the grades. The November 2018 exam was the only assessment that had a significant p-value, but the beta coefficient was 0.003. Instead of contradicting it, this just confirmed the null hypothesis that the true value of the beta coefficient was indeed very close to zero, and that the minutes watched only had a very slight impact on the grades in this case.

As argued in Section 1, attempting to measure the effect of video watching on grades directly is not ideal; and in our case, no relationship between minutes watched and student performance could be found.

When it comes to observational studies that have inherent bias, the question should not be, "Do the videos cause significantly better performance in the treatment group than in the control group?", but rather, "Do the videos cause a significant positive *change* (or improvement) in grades for the treatment group compared with the control group?" In observational studies the aim should be to look at changes in the treatment group's grades, not at the actual grades, and to compare these with changes in the control group's grades. For this reason, the next section focuses on this comparison.

### 4.3 Comparison between the mean grades for the video and no-video groups

Figures 4 and 5 display the means of the video (treatment) and no-video (control) groups for assessments in 2017 and 2018 respectively. Figure 4 shows that the treatments group consistently had higher average grades than the control group for the 2017 assessments.

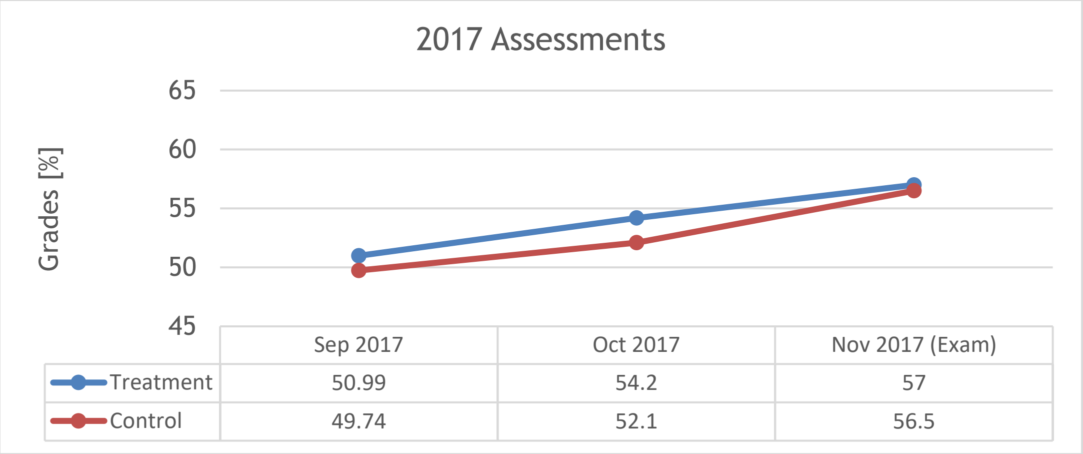
For the 2018 assessments, a third group, called 'treatment 2$^{nd}$', was identified. These were students who did not watch videos in the first semester but started watching in the second semester. They were therefore part of the control group for Semester 1 and part of the treatment group for Semester 2.

In Figure 5, the treatment group consistently got better marks than the control group. However, as discussed in Section 1, because this is an observational study, the difference in marks could be due to confounding variables, such as student motivation.

The treatment 2$^{nd}$ group, however, showed a dip below the control group for the June 2018 exam. After this, their marks started following the treatment group's marks when they started watching videos and were higher than the control group. The exception was the November 2018 exam, for which the average marks for the treatment 2$^{nd}$ group were the same as those for the control group.

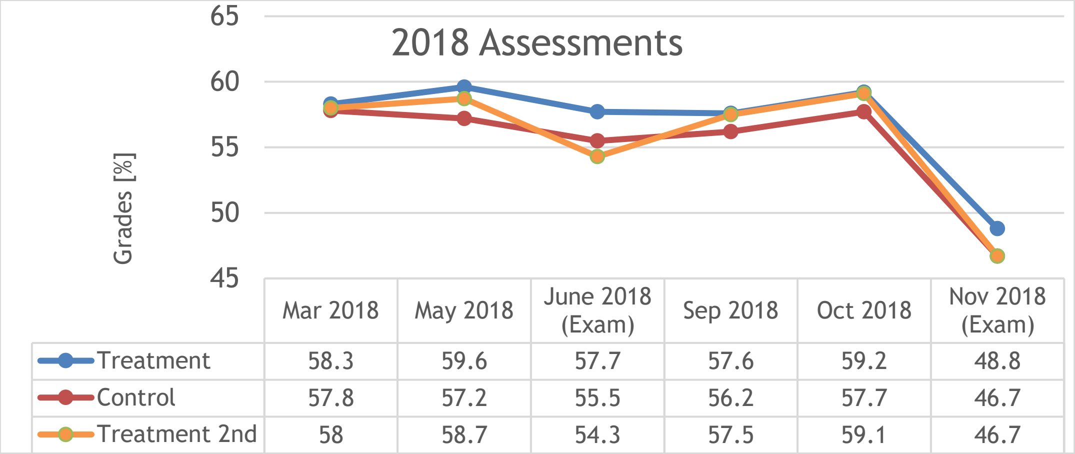
The dip of the treatment 2$^{nd}$ below the Control group may be an instance of Ashenfelter's dip [28]: their drop in grades for the June exam might be why they began watching videos in the second semester.

To investigate the differences in grades further, two-sided t-tests were used to calculate p-values. These were used to assess whether the differences in grades between the video and the no-video groups were significant. For most of the assessments, the p-values were higher than α = 0.05, indicating that there were no significant differences. For the October 2017 test, the means of the video group were significantly higher ($p$ = 0.016) than those of the no-video group. The reason for this significance may be the short period — only one month —between the September 2017 test and the October 2017 test. This suggests that the test content and the 19 videos created for one chapter were very focused. There was thus a low variance in the range of material covered, and perhaps this low variance made the videos effective.



| | Sep 2017 | Oct 2017 | Nov 2017 (Exam) |
|---|---|---|---|
| Treatment | 50.99 | 54.2 | 57 |
| Control | 49.74 | 52.1 | 56.5 |

**Figure 4: Average grades for control and treatment groups for Semester 2, 2017**



| | Mar 2018 | May 2018 | June 2018 (Exam) | Sep 2018 | Oct 2018 | Nov 2018 (Exam) |
|---|---|---|---|---|---|---|
| Treatment | 58.3 | 59.6 | 57.7 | 57.6 | 59.2 | 48.8 |
| Control | 57.8 | 57.2 | 55.5 | 56.2 | 57.7 | 46.7 |
| Treatment 2nd | 58 | 58.7 | 54.3 | 57.5 | 59.1 | 46.7 |

**Figure 5: Average grades for control and treatment groups for Semester 2, 2018**

### 4.4 DID values based on initial grades

In this section, the DID values to estimate treatment effects were computed for the periods from mid-year to the November exam for both 2017 and 2018. First, the DID value for the entire cohort, broken up into the treatment and control groups, was computed for both years. Next, the DID values for passing and failing students, based on their initial mid-year grades, were computed. This was done to investigate how the initial grades of the students at mid-year would affect (1) how they performed in the second semester, and (2) whether the use of videos was effective. We therefore controlled for mid-year passing or failing grades and estimated the DID values.

Table 2 presents the results for the period from mid-year to the November exam for both 2017 and 2018. Although not significant at $\alpha = 0.05$, the failing group for 2017 showed a significant DID estimate of 4.3% at $\alpha = 0.1$, whereas all of the other groups had a non-significant DID estimate.

**Table 2: Changes and DID estimates for the entire cohort and for failing or passing at mid-year, from mid-year to November exam (2017—2018)**

| Year | Mid-year [%] | Change [%] | Change (videos) [%] | Change (no-videos) [%] | DID [%] | p-value |
|---|---|---|---|---|---|---|
| 2017 | Cohort | -1.23 | -1.15 | -1.47 | 0.32 | 0.82 |
| 2017 | Failing | 3.10 | 4.30 | 0.00 | 4.30 | 0.07 |
| 2017 | Passing | -2.30 | -2.40 | -1.90 | -0.50 | 0.70 |
| 2018 | Cohort | -11.03 | -11.30 | -10.60 | -0.70 | 0.59 |
| 2018 | Failing | -5.60 | -5.20 | -6.00 | 0.82 | 0.65 |
| 2018 | Passing | -12.80 | -13.00 | -12.30 | -0.64 | 0.62 |

Finally, the initial mid-year grades were further broken up into control groups, based on the ranges in grades. The aim was to see how different initial ranges of mid-year grades affected the DID values. Table 3 presents the DID values for these sub-groups from mid-year to the November exam for both years. The range <30 had an insignificantly small number of students in both years and was left out of the analysis. For the group that obtained between 40% and 49% at mid-year 2017, significant DID values were computed.

For 2018, this group had a p-value of 0.11, which was lower than any of the other p-values from this year, suggesting that, at a higher significance level, the DID values would be significant. The results suggest that those that were just under the passing grade were most impacted by the videos, and that just missing the 50% mark may significantly motivate students to use the videos intervention. Further study needs to be carried out to determine why this specific sub-group was impacted while a nearby sub-group, such as the 50%-59% group, was not. This will require interviewing students in these sub-groups to obtain further insight.

In 2017, the >80 group also had a significant DID value, if α = 0.1, but not for α = 0.05. This indicates that there could be some benefit for distinction students to keep their grades high by watching the videos. Unlike the 40%-49% group, however, this was not significant in the 2018 group.

**Table 3: Changes and DID estimates based on initial mid-year grades, from mid-year to November exam (2017 — 2018)**

| Year | Mid-year [%] | Change [%] | Change (videos) [%] | Change (no-videos) [%] | DID [%] | p-value |
|---|---|---|---|---|---|---|
| 2017 | >80 | -5.78 | -1.75 | -15.85 | 14.10 | 0.07 |
| 2017 | 70-79 | -1.72 | -2.47 | 0.45 | -2.92 | 0.22 |
| 2017 | 60-69 | -3.35 | -3.00 | -4.56 | 1.56 | 0.44 |
| 2017 | 50-59 | -1.63 | -2.05 | -0.20 | -1.85 | 0.27 |
| 2017 | 40-49 | 2.62 | 4.27 | -1.40 | 5.67 | 0.02 |
| 2017 | 30-39 | 2.60 | 0.74 | 8.73 | -7.99 | 0.13 |
| 2018 | >80 | -16.50 | -17.00 | -15.30 | -1.75 | 0.65 |
| 2018 | 70-79 | -15.70 | -15.80 | -15.60 | -0.20 | 0.92 |
| 2018 | 60-69 | -12.60 | -13.30 | -11.70 | -1.60 | 0.30 |
| 2018 | 50-59 | -10.80 | -10.70 | -11.10 | 0.34 | 0.80 |
| 2018 | 40-49 | -7.00 | -5.90 | -8.80 | 2.90 | 0.11 |
| 2018 | 30-39 | -2.60 | -4.30 | -1.20 | -3.11 | 0.39 |

## 5 DISCUSSION

Although the direct comparison of grades between those who watched videos and those who did not found no association except for one assessment (the Oct 2017 test), the DID values did find treatment effects. A main finding in this study, using the DID estimations, is that the initial grade is a predictor of the treatment effect for certain percentage ranges — or, put the other way around, the treatment effect is dependent on the initial grade; and that, for those initially failing, the intervention is found to be effective. The literature has linked initial factors to performance, but none of the literature has found that initial factors link to the treatment effect. This finding suggests that those who are failing (and specifically those in the 40%-49% range) use the videos in the most productive way. A reason for the videos being effective for this group might lie more in the student than in the videos. For example, these students might use the videos in a targeted or strategic way [7], meaning that they use them to focus on topics with which they are struggling.

Another reason for the videos' effectiveness in this group might be intrinsic motivation [11]. Those who initially perform poorly may have a strong incentive to use the available resources as effectively as possible. Therefore, although the intervention is seen as a valuable resource, the motivation of the student is seen to be the source of the videos' effectiveness.

This study has perhaps found a proxy for motivation: the initial grade that a student obtains early in a semester. This study suggests that an intervention has no value in itself, but that it needs to be combined with other factors within the student. A similar finding by Boaler *et al.* showed that mindset beliefs impact academic performance [16].

Although future interventions will always be open to any students, this finding suggests that the focus needs to be placed on the group of students who are just under the failing grade, encouraging them to use the videos as much as possible. The idea then is that this practice would improve the pass rate in the course.

## 6 CONCLUSION

This study proposed that a reason for the mixed results in student performance reported in the literature, but often overlooked, is the use of the wrong measurement techniques to measure treatment effects in observational studies. To improve interventions, it is crucial to measure their effectiveness accurately. A potential reason provided in this research for mixed results in the effect of videos on grades is that the majority of the studies are observational (students self-electing to watch the videos instead of being placed in randomised control groups). Thus, the literature is mostly measuring video effectiveness without considering confounding factors (e.g, students being more motivated or diligent owing to an awareness of their failing grades, and so taking the initiative to make a greater effort, including watching videos).

In the current observational study, carried out over two years, we also did not find any association between videos and performance when comparing grades with the amount of time spent watching videos. The difference-in-differences (DID) method was then applied to measure more accurately the treatment effects on changes in grades. DID estimates were measured (1) for the entire cohort; (2) based on initial grades; and (3) based on the percentage of videos watched. The findings were that (1) there was no significant impact on the overall group; and (2) the grades at the beginning of the time span were a predictor of the videos' impact — and, more specifically, that the sub-group with initial failing grades of between 40% and 49% was significantly impacted by the videos.

## 7 FURTHER WORK

Intrinsic factors within a student — such as their motivation and their method of video usage — are seen as major factors in how effective the video intervention is. Therefore, in follow-up studies, the entire cohort of students could be interviewed and ask various questions about their motivational state and their method of usage. This opens an important window into combining actual watch time, academic performance, and the motivational state of the students, to establish the differences between cohorts.

A more thorough study of the quality and objective of the videos, especially after the start of COVID-19, in combination with students' views on the videos, could also provide a better understanding of how to improve performance. This could also be used to investigate how COVID-19 has impacted student behaviour and whether there is a difference in the impact of video watching on student performance. If the objective of the videos focuses more on laying the foundations of concepts, it makes sense that failing students would benefit more from the videos, while distinction students would benefit by spending more of their time on mastering more complicated topics not addressed in the videos. It is believed that the students themselves play an integral part in moving the conversation forward about using videos as an online learning tool.

The amount of video watch time could also benefit from further in-depth analysis and from more studies of where the actual minutes were spent, whether the videos were watched to the end or repeatedly, and the students' attention span of while watching the videos.

Finally, although global (group and sub-group) analyses have some value, the literature increasingly shows the need to measure individualised treatment effects that are shown to be more effective for each student [9, 29, 30]. Therefore, further research could focus on individualised treatment effects that could assist the students with personalised learning [31].

## REFERENCES

[1] F.J. Boster, G.S. Meyer, A.J. Roberto, C. Inge and R. Strom, "Some effects of video streaming on educational achievement," *Communication Education,* vol. 55, no. 1, pp. 46-62, 2006.

[2] M. Wieling and W. Hofman, "The impact of online video lecture recordings and automated feedback on student performance," *Computers & Education,* vol. 54, no. 4, pp. 992-998, 2010.

[3] A. Williams, E. Birch and P. Hancock, "The impact of online lecture recording on student performance," *Australasian Journal of Educational Technology,* vol. 28, no. 2, pp. 1-9, 2012.

[4] T. Traphagan, J.V. Kucsera and K. Kishi, "Impact of class lecture webcasting on attendance and learning," *Educational Technology Research and Development, vol.* 58, no. 1, pp. 19-37, 2010.

[5] A. Le, S. Joordens, S. Chrysostomou and R. Grinnell, "Online lecture accessibility and its influence on performance in skills-based courses," *Computers & Education,* vol. 55, no. 1, pp. 313-319, 2010.

[6] J.A. Brotherton and G.D. Abowd, "Lessons learned from eClass," *ACM Transactions on Computer-Human Interaction,* vol. 11, no. 2, pp. 121-155, 2004.

[7] W. Leadbeater, T. Shuttleworth, J. Couperthwaite and K.P. Nightingale, "Evaluating the use and impact of lecture recording in undergraduates: Evidence for distinct approaches by different groups of students," *Computers & Education,* vol. 61, no. 1, pp. 185-192, 2013.

[8] D. Faraoni and S.T. Schaefer. "Randomized controlled trials vs. observational studies: Why not just live together?" *BMC Anesthesiology*, vol. 16, no. 1, 2016, pp. 1-4.

[9] J. Beemer, K. Spoon, L. He, J. Fan, and R.A. Levine, "Ensemble learning for estimating individualized treatment effects in student success studies," *International Journal of Artificial Intelligence in Education,* vol. 28, no. 3, 2017, pp. 315-335.

[10] J.J. Heckman. "Selection bias and self-selection," in S.N. Durlauf and L.E. Blume, eds, Econometrics, London, Palgrave Macmillan, pp. 201-224, 1990.

[11] D. Muller, J. Eklund and M.D. Sharma, "The future of multimedia learning: Essential issues for research," in Paper presented at the Australian Association for Research in Education, Sydney, vol. 18, 2005.

[12] M. Ibrahim, "Implications of designing instructional video using cognitive theory of multimedia learning," *Critical Questions in Education*, vol. 3, no. 2, pp. 83-104, 2016.

[13] A. Hansch, L. Hillers, K. McConachie, C. Newman, T. Schildhauer and J.P. Schmidt, "Video and online learning: Critical reflections and findings from the field," *HIIG Discussion Paper Series*, Berlin, Alexander von Humboldt Institute for Internet and Society, 2015.

[14] V. Kettanurak, K. Ramamurthy and W. Haseman, "User attitude as a mediator of learning performance improvement in an interactive multimedia environment: An empirical investigation of the degree of interactivity and learning styles," *International Journal of Human-Computer Studies,* vol. 54, no. 4, pp. 541-583, 2001.

[15] H. Astleitner and C. Wiesner, "An integrated model of multimedia learning and motivation," *Journal of Educational Multimedia and Hypermedia,* vol. 13, no. 1, pp. 3-21, 2004.

[16] J. Boaler, J.A. Dieckmann, G. Pérez-Núñez, K.L. Sun and C. Williams, "Changing student minds and achievement in mathematics: The impact of a free online student course," *Frontiers in Education,* vol. 3, no. 26, pp. 1-7, 2018.

[17] C.M. Chen and C.H. Wu, "Effects of different video lecture types on sustained attention, emotion, cognitive load, and learning performance," *Computers & Education* vol. 80, no.1, pp. 108-121, 2015.

[18] P.C. Austin, "An introduction to propensity score methods for reducing the effects of confounding in observational studies," *Multivariate Behavioral Research,* vol. 46, no. 3, pp. 399-424, 2011.

[19] A.M. Ryan, E. Kontopantelis, A. Linden, and J.F. Burgess, Jr. "Now trending: Coping with non-parallel trends in difference-in-differences analysis," *Statistical Methods in Medical Research,* vol. 28, no. 12, pp. 3697-3711, 2019.

[20] A. Rambachan and J. Roth, "An honest approach to parallel trends," Working paper, Harvard University, 2019. [Online]. Available: https://scholar.harvard.edu/files/jroth/files/honestparalleltrends_main.pdf [Accessed June 6, 2021].

[21] D. Card and A.B. Krueger, "Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania," *The American Economic Review*, vol. 84, no. 4, pp. 772-793, 1994.

[22] O. Ashenfelter and D. Card, "Using the longitudinal structure of earnings to estimate the effect of training programs," *The Review of Economics and Statistics*, vol. 67, no. 4, pp. 648–660, 1985.

[23] B.D. Meyer, W.K. Viscusi and D.L. Durbin, "Workers' compensation and injury duration: Evidence from a natural experiment," *The American Economic Review*, vol. 85, no. 3, pp. 322-340, 1995.

[24] T. Conley and C. Taber, "Inference with 'difference in differences' with a small number of policy changes," *National Bureau of Economic Research*, vol. 93, no. 1, pp. 113-125, 2011.

[25] M. Bertrand, E. Duflo and S. Mullainathan, "How much should we trust differences-in-differences estimates?" *The Quarterly Journal of Economics*, vol. 119, no. 1, pp. 249-275, 2004.

[26] J. Gareth, D. Witten, T. Hastie and R. Tibshirani, *An introduction to statistical learning with applications in R*, New York, Springer, 2013.

[27] Panopto$^{TM}$, Panopto: Video made easy, [Online]. Available: https://www.panopto.com [Accessed June 6, 2021].

[28] O. Ashenfelter, "Estimating the effect of training programs on earnings," *The Review of Economics and Statistics*, vol. 60, no. 1, pp. 47-57, 1978.

[29] J. Beemer, K. Spoon, J. Fan, J. Stronach, J.P. Frazee, A.J. Bohonak and R.A. Levine, "Assessing instructional modalities: Individualized treatment effects for personalized learning," *Journal of Statistics Education,* vol. 26, no. 1, pp. 31-39, 2018.

[30] B.I. Smith, C. Chimedza and J.H. Bührmann, "Global and individual treatment effects using machine learning methods," *International Journal of Artificial Intelligence in Education*, vol. 30, no. 1, pp. 431- 458, 2020.

[31] B.I. Smith, C. Chimedza and J.H. Bührmann, "Individualized help for at-risk students using model-agnostic and counterfactual explanations," *Education and Information Technologies*, Online first articles, 2021. [Online]. https://link.springer.com/article/10.1007%2Fs10639-021-10661-6 [Accessed July 23, 2021].