Examining the Effect of Virtual School Enrollment on Student Outcomes

Nathan Calverley*

November 18, 2013

Abstract

This study analyzes the effects of full-time virtual school enrollment on the likelihood of high school graduation and post-secondary enrollment in Wisconsin. It also explores the impact of full-time virtual school enrollment on student standardized test scores. The study uses propensity score matching to produce estimates of the effects of virtual school enrollment, and finds that students who enroll in virtual schools are 6-22% less likely to graduate from high school, 4-19% less likely to enroll in a post-secondary institution, and experience declines in mathematics test performance of -0.19 standard deviations. Implications for the current state of full-time virtual school policy are discussed.

1 Introduction

As the world has become more connected over the last couple of decades, technology has become increasingly pervasive in nearly all aspects of society. One of the fastest growing areas in which technology is being leveraged is education, as schools of all types have dramatically increased their use of technology to educate kids. Rural schools in particular have reported large increases in the use of virtual education, with 46% of rural districts reporting student participation in distance education as of 2005, compared with only 23% in suburban districts (Setzer and Lewis, 2005).

The education community seems to be accepting that virtual schooling is quickly becoming a part of the mainstream, with more and more people viewing distance learning as a legitimate method of delivering quality education. In June of 2013, a Boston-based research and consulting firm published the results of a survey it conducted of college students in the United States on their opinions towards virtual education, and the results were striking: 43% of students believed that virtual education could provide them with at least as good of an education as traditional in-class studies. However, despite this confidence in the quality of virtual schooling, the vast majority of those same students (78%)

*Nathan Calverley is a Research Analyst at the Wisconsin Department of Public Instruction. The views expressed in this paper do not necessarily represent the positions or policies of the Wisconsin Department of Public Instruction. No official endorsement by the Wisconsin Department of Public Instruction is intended, nor should one be inferred.

believed that it was easier to learn in the traditional classroom environment (Millennial Branding, 2013). But is this actually the case?

A major problem with the expansion of full-time distance education is that quantitative research on virtual schools has failed to keep up with their rapid expansion. Quality research in the K-12 realm is particularly elusive, with only a handful of studies having been published over the last decade. In an attempt to add to the (currently shallow) pool of K-12 virtual research, this study analyzes a sample of Wisconsin students who enrolled full-time in virtual schools during their high school careers, and attempts to determine what, if any, impact this enrollment had on a student's probability of graduating from high school and attending a post-secondary institution. It will also examine the effect of full-time virtual school enrollment on standardized test scores. Finally, it will offer some insights into future research on the topic of distance education.

2 Virtual Schools: Existing Research

There is an alarming shortage of quality research on the long-term academic outcomes of virtual school students. In fact, a detailed search of academic databases failed to yield any studies connecting virtual school enrollment to graduation and post-secondary outcomes. There is also a shortage of quality research on the short-term academic outcomes of virtual school students. A 2010 study by the U.S. Department of Education attempting to conduct a meta-analysis of publications on the effect of virtual schooling on K-12 education yielded only nine studies published between 1996 and 2008 that the authors found acceptable for inclusion in the analysis. However, the few studies that do exist on virtual K-12 education offer valuable insights into virtual schools and how they differ from traditional schools.

Studies on the effects of K-12 virtual schooling generally find that students enrolled in virtual courses typically perform at levels comparable to students who take the same courses in a face-to-face environment, a finding which has recently taken on an almost pejorative connotation in virtual education literature and become known as the ”no significant difference” phenomenon. For example, an oft-cited meta-analysis of 19 studies on K-12 distance education by Cavanaugh (2001) found no significant overall effect size for virtual education. However, this study has received criticism from others based on the fact that only 10% of the studies in Cavanaugh's meta analysis used an experimental or quasi-experimental approach (O'Dwyer, Carey, and Kleiman, 2007). A meta-analysis conducted by the U.S. Department of Education (2010) identified 50 independent effects and concluded an overall positive effect of virtual education on student performance outcomes. However, the study warned that so few of these effects were from K-12 studies that researchers should be cautious in generalizing the result to K-12 learning.

Other criticisms of K-12 meta-analyses is that they ignore important marginal effects found in some of their studies, obfuscating important findings through aggregation to a mean effect of zero. For example, a meta analysis conducted by Zhao et al. (2005) found that studies finding positive effects for virtual education shared a common set of characteristics that were absent from studies finding negative effects, including large amounts of interaction between students and instructors, the presence of a live human instructor (rather than simple interaction with a computer), and the characteristics of the students taking the course. After all, if one group of students experiences positive outcomes, and one group of students experiences negative outcomes, it is not particularly useful to conclude that overall, nothing happened.

Other studies have tried to highlight the importance of more intangible outcomes for virtual students when finding no significant effect on achievement outcomes. For example, O'Dwyer, Carey, and Kleiman (2007) examined the Louisiana Algebra I online course, a state-sponsored online course targeted to students in rural or urban areas lacking certified algebra teachers, found that students who took algebra I online and students who took it in a face-to-face environment had no statistical difference in performance on a pre- and post-course assessment. However, they found that, when posed with the question of, "Did the course provide a good learning experience?", only 38.5% of the virtual students responded "yes", compared with 62.8% of the face-to-face students, despite actual performance levels of the two groups being statistically similar.

Rockman et al (2007) examined the outcomes of students in a virtual Spanish class in West Virginia, finding that students in the virtual classes performed at levels similar to face-to-face students on the multiple-choice section of the state's Spanish assessment, but significantly lower on the oral fluency and writing sections of the assessment. These findings are particularly telling when one considers that the students placed into the virtual Spanish class were typically hand-picked by principals as students who were more motivated, had higher standardized test performance, and were deemed more likely to succeed in the class, yet struggled compared to their peers in traditional classrooms (Rockman et al, 2007, pg. 15).

Other studies have found that the level of interaction between instructors and students, and between the students themselves, is diminished in both quantity and quality in virtual schools. For example, Kozma et. al. (2000) examined a trio of high school courses offered both online and in a physical classroom, and found no statistical difference in the performance of students on a key course assignment given to both groups. However, they also found that student interaction with instructors and with each other was significantly diminished in the virtual classes: "This lack of frequency, depth, and connectedness in student interaction...is likely to have negatively affected the amount [students] learned." (Kozma et al, 2000)

More recent research, however, is beginning to leave the "no significant difference" paradigm behind as quantitative studies of higher quality have begun to emerge. A 2011 study by the Center for Research on Education Outcomes (CREDO) at Stanford University compared the academic growth of Pennsylvania students in virtual charter schools to that of a matched sample of similar students in traditional schools, finding that students in the virtual charters grew at significantly lower rates than their peers in traditional schools in both mathematics and reading (CREDO, 2011).

Similarly, a 2013 study by the National Education Policy Center (NEPC) analyzed the academic performance of virtual schools across the entire United States. By compiling school rating designations for all virtual schools, the authors determined that in the 2011-12 school year, nearly 72% of virtual schools with a rating (228 schools in total) were designated as "academically unacceptable". Further, by examining four-year graduation rates, the authors determined that the overall on-time graduation rate for virtual schools was a meager 37.6%, compared with the national average of 79.4% (Molnar et al., 2013).

These figures are naive estimates in that they were not derived from an experimental or quasi-experimental design, but they do paint a sordid picture of the state of virtual schools across the U.S.

Overall, there is much disagreement in the sparse literature on the effects of K-12 virtual education vis-a-vis traditional face-to-face education, with many studies reaching a conclusion of ”no significant difference”. However, recent research that is more generalizable to K-12 outcomes is beginning to question the ”no significant difference” phenomenon, and has become increasingly critical of the performance of virtual schools.

3 Purpose of Study

Lawmakers and researchers have often complained that the expansion of virtual schools in Wisconsin over the last several years has been premature, citing the lack of available studies of virtual schools and their impact on K-12 student performance. Wisconin State Representative Steve Kestell, chairman of the state Assembly Committee on Education, has expressed discomfort with the current state of knowledge surrounding virtual schools (Litke, 2012), and the National Education Policy Center released a report in 2013 calling for a nationwide moratorium on the growth of full-time online schools until more effective and substantive research has been conducted (Molnar et al., 2013).

This study attempts to help alleviate the dearth of research on virtual schools by analyzing the effects of *full-time* virtual school enrollment on several key academic outcomes. Full-time enrollment means that students take their entire course load in a virutal environment, as opposed to supplementary virtual schooling in which students take one or two courses online, with the rest of their coursework being done in a traditional brick and mortar school.

This study uses a quasi-expirimental design to answer the following questions:

1. *What is the difference in the likelihood of graduation for students who attend a full-time virtual school for at least one year, compared with similar students who do not?*

2. *What is the difference in the likelihood of post-secondary enrollment for students who attend a full-time virtual school for at least one year, compared with similar students who do not?*

3. *What is the effect of full-time virtual school enrollment on standardized test performance?*

The rest of the paper proceeds as follows: Section 4 of the study will discuss the analytical approach taken in attempting to answer the research questions, while section 5 will go into detail about the data used in the study. Section 6 will provide an in-depth discussion on the strength of the analytical approach used, and Sections 7 and 8 present the results of the analyses. Finally, Section 9 provides a discussion of the results and implications for policy and future research.

4 Theoretical Framework

Firstly, it is necessary to address some key nomenclature that will be used throughout the rest of this study. The notion of receiving ”treatment” refers to any student that experiences the event of interest - in this case, enrolling full-time in a virtual school at any point during high school.

The fundamental problem that must be overcome in any program evaluation is the absence of alternative outcome data. We can observe the outcomes of students who receive treatment and the outcomes of those who do not, but we can never observe the outcomes of students who received treatment in a world where they instead did not, and vice-versa for non-treated students. In other words, it is impossible for us to know what would have happened to students who attended a virtual school during high school, had they instead attended only traditional brick-and-mortar schools. Fortunately, there are statistical methods for overcoming this problem.

Propensity Score Matching

This analysis utilizes propensity score matching in order to derive an estimate of the effect of treatment on student outcomes. Propensity score matching compares similar students to each other based on their probability of receiving treatment; this probability is the propensity score. Once each student’s propensity score is estimated, the outcomes of students with similar propensity scores but different treatment statuses are compared. This allows us to produce an estimate of the effect of treatment on student outcomes.

The propensity score can be defined theoretically as such:

$$P(x) = Pr(D = 1|X = x)$$

Where P(x) is the propensity score, or the probability that a student receives treatment (D = 1) conditional on a vector of characteristics, X. Once the propensity score is estimated, the treatment and control groups are created by matching treated and untreated students with similar propensity scores. After the treatment and control groups have been assembled, the average treatment effect (ATE) can be estimated.

The biggest advantage to using propensity score matching over other more traditional statistical techniques (such as OLS regression modeling) is that propensity score matching is a non-parametric measure, meaning that assumptions about the functional form of the model are relaxed. While OLS requires a linear relationship in the data in order for the model to produce estimates of the effect of treatment that are unbiased, propensity score matching does not have this requirement.

The estimation of the propensity score does, however, carry some very important assumptions of its own. First, in order for estimates of the ATE to be unbiased, the assumption is made that outcomes for students who do not receive treatment are independent of the outcomes for those same students had they instead received treatment. In other words, within subgroups defined by a vector of variables, X, your probability of graduating high school without attending a virtual school is unrelated to your probability of graduating high school if you did enroll in a virtual school. This assumption is often refered to as the conditional independence or strong ignorability assumption.

The second requirement of propensity score matching is an area of common support - the area in which the distribution of Pr(D = 1 | X = x, D = 1) overlaps Pr(D = 1 | X = x, D = 0). In other words, the area of common support is the range of propensity score values in which there are both treated and untreated observations.

The final, and often overlooked, requirement in order for propensity scores to yield unbiased estimates of the ATE is sufficient pre-treatment covariate balance. The goal of the propensity score is to balance the pre-treatment characteristics of the treatment and control groups (Steiner, 2011), as this is a requirement for (though not a guarentee of) the satisfaction of the conditional independence assumption. The extent to which the specification of the propensity score was successful can be measured by comparing the pre-treatment characteristics of the treatment and non-treatment groups and assessing the degree to which these characteristics differ statistically (Pattanayak, Rubin, and Zell, 2011). When covariate balance cannot be achieved, however, there are methods for estimating the ATE that adjust for bias resulting from covariate imbalance. This study will implement several methods of estimation using the propensity score, the details of which will be discussed further in Section 6.

5 Data

Since the 2006-07 school year, full-time virtual school enrollment in Wisconsin has increased by a factor of four. Distance learning has become more and more popular, especially among students in the high school grades, who comprise the majority of students in virtual schools. Until the 2012-13 school year, a state-wide cap was in place that limited the number of students that could enroll in virtual schools through open enrollment to 5,250 students, with no cap on within-district enrollment. However, changes to state law in the 2011-13 biennial state budget removed that cap, and as a result, virtual school enrollment increased by 37% in the 2012-13 school year.

Table 1: Counts by Student Group in Overall Sample, Treatment vs. Non-Treatment

	Students in Treatment	Percent	Students in Non-Treatment	Percent	Total	Percent
All	3,031	100%	206,209	100%	209,240	100%
Asian	35	1.2%	7,551	3.7%	7,586	3.6%
Black	227	7.5%	20,379	9.9%	20,606	9.8%
Hispanic	168	5.5%	12,958	6.3%	13,126	6.3%
Am. Indian	65	2.1%	3,067	1.5%	3,132	1.5%
White	2,536	83.7%	162,254	78.7%	164,790	78.8%
English Learners	23	0.8%	6,443	3.1%	6,466	3.1%
Students with Disab.	326	10.8%	26,880	13.0%	27,206	13.0%
Econ. Dis.	936	30.9%	64,729	31.4%	65,665	31.4%

The student-level data used in this analysis come from Wisconsin's Longitudinal Data System (LDS), which is maintained by the state Department of Public Instruction (DPI). The data elements used in the study include:

- Wisconsin Knowledge and Concepts Examinations (WKCE) performance: The WKCE is Wisconsin's state-administered test that is given in the fall of every year to students in grades three through eight, and also to tenth graders. In order to allow for WKCE scores to be analyzed across grades and years, the scores were normalized, and should thus be interpreted in units of standard deviations;

- Attendance rates and school information: Attendance data is collected to the half-day, and is aggregated at the student level across all schools attended in each school year;

- Disciplinary Data: Information regarding the number of disciplinary incidents involving each student, and the number of days the student was removed from school (out of school suspensions and expulsions) as a result of those incidents, is collected at the end of each year;

- Demographic Data: Information regarding each student's race/ethnicity, age, disability status, FRL eligibility, and English proficiency level is collected at both the start and end of each school year;

- National Student Clearinghouse post-secondary enrollment information: The National Student Clearinghouse collects enrollment data from over 3,300 post-secondary institutions, and provides post-secondary enrollment data each year for students who graduate from Wisconsin high schools and continue on to higher education;

- Open Enrollment Data: The study also utilized open enrollment data from the School Management Services team at DPI, which are collected between February and April each year for the upcoming school year. Open enrollment data contain the names, birth dates, district of residence, and district of open enrollment for each student accepted into an open enrollment school.

Table 2: Counts of Students in Virtual Schools by Lowest Grade Level of Attendance

	Number of Students	Percent
9th Grade	802	26.5%
10th Grade	575	19.0%
11th Grade	909	30.0%
12th Grade	712	23.5%
Total	3,031	100%

Table 1 describes the sample of students used in this analysis. In order to be included in the analysis, students had to be enrolled in Wisconsin public schools for at least three years, and they must have an observed graduation outcome. As Table 1 shows, attendance in virtual schools is generally uncommon. White students in virtual school populations tend to be overrepresented compared to the untreated students, while Asian, Black, and Hispanic students tend to be underrepresented. Extremely few English learners were found in the virtual school sample, but the proportion of virtual schoolers who were disabled or economically disadvantaged was similar to the proportions found in the untreated population. Table 2 summarizes the high school grades in which students *first* attended virtual schools. The data indicate that students have a similar likelihood of enrolling in each of the high school grades, with the largest numbers enrolling in the junior year (30.0%) and the freshman year (26.5%).

6 Specification of the Propensity Score and Estimation Strategy

The propensity score was constructed using logistic regression. Propensity scores were generated on a wide range of pre-treatment indicators, including student demographics, reading and mathematics standardized test scores, attendance rate, age at time of treatment, number of disciplinary incidents, number of days removed from school due to disciplinary problems, school-level characteristics such as the concentration of economic disadvantagedness and school violence, and a range of higher-order and interaction terms implemented to achieve covariate balance. The exact specification of the propensity score can be found in the appendix.

After the propensity score was estimated, three matched data sets were produced to be used in three different estimation strategies:

- Estimation through regression;
- Estimation through inverse probability weighting;
- Estimation through differencing means of a balanced data set.

Estimation through Regression

The strategy for estimating the ATE through regression is to construct a control sample by matching from the pool of possible control students, estimating the outcome of the treatment group conditional on the propensity score under both treatment statuses, and then differencing the conditional outcomes of the treatment group. Mathematically, this is calculated as:

$$A\hat{T}E = \frac{1}{N}\sum_{i=1}^{N}(\hat{Y}_1(e(X_i)) - \hat{Y}_0(e(X_i)))$$

where

$$e(X_i)$$

is the propensity score. Assuming that the propensity score is correctly specified, this method is likely to be effective in eliminating bias resulting from imbalanced covariates (Imbens, 2004). The control group constructed for this estimation method utilized the specification of the propensity score described previously, and summary and balance statistics of the sample can be found in Appendix C.

Estimation through Inverse Probability Weighting

A second strategy used for estimating the ATE when the distributions of pre-treatment covariates differ between the treatment and control groups is a method of weighting observations by their conditional probability of treatment, a method known as inverse probability weighting. The estimator is derived through a weighted least squares regression specified as:

$$Y_i = \alpha + \beta D_i + \varepsilon$$

and with the weights equal to:

$$\lambda_i = \sqrt{\frac{D_i}{e(X_i)} + \frac{1 - D_i}{1 - e(X_i)}}$$

The weighting reduces the bias caused by covariate imbalance by attempting to achieve statistical independence between the covariates and the treatment status (Imbens, 2004). The control group constructed for this estimation method utilized the specification of the propensity score described previously, and summary and balance statistics of the sample can be found in Appendix C.

Estimation through Differencing Means of a Balanced Data Set

The final, and prefered, estimation strategy is through use of a matched data set that has fully balanced pre-treatment covariates between the treatment and control groups. If the propensity score matching has resulted in a sample in which the treatment and control groups have distributions of pre-treatment covariates that are approximately equal, then covariate balance has been achieved and it is possible to derive an estimation of the ATE that is unbiased by simply taking a difference in means of the two groups (Dahejia and Wahba, 1997).

After the specification of the propensity score, two matched data sets were constructed using the nearest neighbor matching algorithm and applying a caliper to ensure that each match was of sufficient quality. This study utilizes matching *with replacement*, meaning that multiple treated observations may be matched to the same untreated observation. The nearest neighbor method matches each treated observation with one untreated observation having the propensity score that most closely matches that of the treated observation, while the caliper ensures that a match can only be valid if the difference between the treated and untreated propensity scores falls within an acceptable range. In other words, if a treated observation's closest match is an untreated observation with a propensity score that deviates from its own by an amount larger than the specified caliper, this match is not considered valid because the two are too dissimilar, even though the matched untreated observation is the closest match available. The caliper used in this study is 0.01.

The caliper of 0.01 was chosen so as to produce a highly balanced data set while preserving the largest number of observations. The size of the caliper suggests that there is very little differentiation in the propensity scores, a result of the fact that it is actually very difficult to predict which students will attend virtual schools in Wisconsin. Other than virtual school students being slightly more white on average than the overall statewide population, virtual school students tend to be quite average on all other performance metrics (Section 6 will explore these metrics futher.) The end result is that the resulting propensity scores tend to be highly clustered. This is not problematic, however, as the goal of the propensity scores is simply to produce a data set wherein the treatment and control groups have balanced pre-treatment covariates (Steiner, 2011).

Table 3: Student Counts and Percents by Subgroup, Matched Data Sample for Estimation through Balanced Data

	Number of Students: Treatment	Percent	Number of Students: Non-Treatment	Percent
All	1,185	100.0%	1,185	100.0%
Asian	15	1.3%	16	1.4%
Black	49	4.1%	64	5.4%
Hispanic	44	3.7%	67	5.7%
Am. Indian	29	2.4%	12	1.0%
White	1,048	88.4%	1,026	86.6%
English Learners	8	0.7%	29	2.4%
Students with Disab.	149	12.6%	142	12.0%
Econ. Dis.	411	34.7%	462	39.0%

Table 3 shows the summary statistics for the data set produced for the estimation of the ATE through differencing means of a balanced data set. The treatment and control groups within the sample closely mirror the characteristics of the overall virtual student population from Table 1. Students in both groups are predominantly white, with the next largest group being economically disadvantaged students. Very few of the students in either group are English learners, and students from all non-white ethnicities are underrepresented compared to state-wide populations.

Table 4: Assessing Covariate Balance of the Sample for Estimation through Differencing Means of a Balanced Data Set

	Mean Treatment	Mean Non-Treatment	P-Value
Reading Score	-0.02	-0.09	0.08
Math Score	-0.17	-0.24	0.05
Attendance Rate	87.12%	87.40%	0.60
Disciplinary Incidents	0.31	0.30	0.86
Days Removed from School	1.35	0.88	0.10
Percent White	0.88	0.87	0.17
Percent Econ	0.35	0.39	0.03
Percent Disab	0.13	0.12	0.66

Assessing the Match Quality

Covariate balance was assessed by stratifying the propensity scores into quintiles, and using simple t-tests to test whether the difference in means of each primary pre-treatment covariate within each strata was statistically spurious. A P-value of 0.05 or less indicates that the difference in means was statistically significant at the 95% level, and suggests that results from the sample may be biased due to statistically significant differences in pre-treatment characteristics between the treatment and non-treatment groups.

Table 5 above shows the results of the balance tests for each matched data set presented *in aggregate* in order to show the general similarity of the treated and matched control group. However, in order to ensure that no imbalances exist along the propensity score distribution, observations were stratified by their propensity scores into quintiles, and the covariates were again tested for balance within each strata. Results of the stratification can be found in Appendix C. The data are largely balanced in each of the five strata, with statistically significant differences in covariate means occuring in only 4 of the 40 observed mean differences.

Establishing Strong Ignorability

The strongest of the underlying assumptions of propensity score matching methods is the assumption of the strong ignorability, as it is impossible to test this assumption empirically. As a result, selection on unobservables becomes even more of a concern to the validity of the matching estimators. Since moving from a traditional school to a virtual school represents a conscious choice on the part of the student and their parents, it is logical to assume that this decision was not made randomly, and was instead the result of a motivation to change the student's academic environment. In an attempt to control for any selection on unobservables, this study utilizes open enrollment data when selecting the control group in each of the matching methods described above.

Wisconsin's open enrollment policy dictates that any student in Wisconsin may apply to enroll in a school outside of their district of residence. In other words, each year, there is a population of students making a conscious decision to change their academic environment from one traditional school to a different traditional school in another district. This population of students provides an excellent counterfactual population. All students in each of the control groups mentioned above were first-time open enrollee students in the treatment year, which should control for any selection on unobservables between the treatment and control groups relating to the decision to change academic environments.

Dependent Variables

As mentioned above, this study had three primary research questions. The dependent variable associated with each of the research questions are as follows:

1. *What is the difference in the likelihood of graduation for students who attend a virtual school for at least one year, compared with similar students who do not?*

   High school graduation is defined as whether or not the student *ever* graduated, as long as graduation occured within six years of entering high school.

2. *What is the difference in the likelihood of post-secondary enrollment for students who attend a virtual school for at least one year, compared with similar students who do not?*

Post-secondary enrollment is defined as whether or not the student *ever* enrolled in a post-secondary institution, regardless of the time between exiting high school and enrolling in a post-secondary institution.

3. *What is the effect of virtual school enrollment on standardized test performance?*

   The dependent variable in this analysis is the student's standardized test scores in math and reading in the first year of enrollment in a virtual school.

7 Results: Effect on Graduation and Post-Secondary Enrollment

Table 5: OLS and Matched Estimates of the Effect of Virtual School Enrollment on Probability of High School Graduation

	OLS Model	Matched Regression Model	Matched IPW Model	Matched Balanced Covariate Model
Estimated ATE	-0.13***	-0.16***	-0.01	-0.17***
Standard Error	0.005	0.015	0.025	0.017
Number of Treated Observations	3,031	1,516	1,516	1,185
Number of Untreated Observations	206,209	866	866	846

What is the difference in the likelihood of graduation for students who attend a virtual school for at least one year, compared with similar students who do not?

Table 5 displays the results of the impact of virtual school attendance on probability of graduation. Standard errors on each of the three matched models are bootstrapped standard errors derived from 1,000 iterations. In all of the statistical models applied to the sample data, virtual school enrollment was associated with a negative effect on the probability of high school graduation. The OLS model found an effect of -13% in the likelihood of graduation associated with virtual school enrollment, which was statistically significant at the 99% level. The matched model using the regression estimation method produced an ATE of -16%, which was statistically significant at the 99% level. In contrast, the matched model using the inverse probability weighting method produced an estimate of only -1%, which was not statistically significant at any level. Finally, the balanced covariate matched model, which is the prefered model, produced an estimate of -17%, which was significant at the 99% level.

The parameter estimate produced by the IPW model warrants further dicussion, as it differs significantly from the other estimates. Recall that the data set produced for the IPW model with the goal of being completely balanced on pre-treatment covariates; rather, the propensity scores would be used in calculating a set of weights to be used in a weighted least squares model, which would adjust for imbalances in the pre-treatment covariates. Also recall that there is very little differentiation in the calculated propensity scores. This lack of differentiation in propensity scores mathematically translates into a lack of differentiation in the weights assigned to each observation, limiting the model's ability to adjust for imbalances in the pre-treatment covariates. If the control group displayed more barriers to graduation in their pre-treatment covariates than the treatment group, then the IPW estimate will suffer from attenuation bias. Examining the pre-treatment characteristics of the treatment and control groups in the IPW sample (found in Appendix C.2), we see that this is indeed the case. In short, the IPW model estimate of the effect of virtual school enrollment on graduation should be interpreted as an extremely conservative estimate.

Table 6: OLS and Propensity Score Matched Estimates of the Effect of Virtual School Enrollment on Probability of Post-Secondary Enrollment

	OLS Model	Matched Regression Model	Matched IPW Model	Matched Balanced Covariate Model
Estimated ATE	-0.13***	-0.08***	-0.03*	-0.10***
Standard Error	0.010	0.013	0.016	0.015
Number of Treated Observations	3,031	1,516	1,516	1,185
Number of Untreated Observations	206,209	866	866	846

What is the difference in the likelihood of post-secondary enrollment for students who attend a virtual school for at least one year, compared with similar students who do not?

Table 6 displays the results of the impact of virtual school attendance on probability of post-secondary enrollment. Standard errors on each of the three matched models are bootstrapped standard errors derived from 1,000 iterations. In all of the statistical models applied to the sample data, virtual school enrollment was associated with a negative effect on the probability of post-secondary enrollment. The OLS model found an effect of -13% in the likelihood of post-secondary enrollment associated with virtual school enrollment, which was statistically significant at the 99% level. The matched model using the regression estimation method produced an ATE of -8%, which was statistically significant at the 99% level. The matched model using the inverse probability weighting method produced an estimate of -3% (with the same caveats applying here as in the previous results regarding the conservativeness of the IPW model), which was significant at the 90% level. Finally, the balanced covariate matched model, which is the prefered model, produced an estimate of -10%, which was significant at the 99% level.

8 Results: Effect on Standardized Test Scores

This section of the analysis focuses on the effect of virtual school enrollment on standardized test scores. As mentioned earlier, in the 2012-13 school year, a statewide cap on virtual school enrollment was lifted, resulting in an increase in statewide virtual school enrollment of just under 1,800 students. This change in the enrollment cap provides a unique opportunity to study the effects of virtual school enrollment on student achievement as measured by the Wisconsin Knowledge and Concepts Examination, the state's assessment of academic achievement used for accountability purposes. The following analysis works on the assumption that students who enrolled in virtual schools for the first time in the 2012-13 school year had intended to enroll in prior years, but were prevented from doing so by the statutory cap on virtual school enrollment.

Analytical Strategy

In order to estimate the effect of virtual school enrollment on standardized test performance, a differences-in-differences analysis was conducted. In the differences-in-differences framework, the analytical structure is defined by two groups of observations, observed at two different periods in time. One group (the treatment group) is exposed to a treatment (in this case, attending a virtual school) in the second time period, but not in the first; the other group (the control group) is not exposed to the treatment in either time period. The average gain in each group's test scores are measured in each period, and then the average gain in the control group is subtracted from the average gain in the treatment group. This differencing removes bias in the groups' scores resulting from trends over time, and well as from any existing differences in group performance.

The differences-in-differences model can be defined as such:

$$Y_i = \alpha + \beta T_i + \gamma t_i + \delta(T_i * t_i) + \varepsilon$$

Where:

β = The treatment group effect, which is *not* the treatment effect, but rather a dummy that accounts for underlying differences between the treatment and non-treatment groups.
γ = A year dummy, which takes on a value of 1 when the school year is 2013 (the treatment year), and 0 when the year is not 2013.
δ = The average treatment effect, the primary outcome measure.

Defining the Treatment and Control Groups

Because differences-in-differences requires trend data, both groups were limited to students who had at least three years of test data. Test scores were normalized in order to allow comparisons across grades and over time. The treatment group consisted of any student who had the required amount of test data, who attended traditional schools from 2010-2012, and who attended a virtual school for the first time in 2013. The control group was constructed from students who had test scores for all years from 2010-2013 and who never attended a virtual school. Following the strategy for controlling for unobservables discussed in the previous section, all students in the control group were first-time open enrollee students in the treatment year.

Table 7 gives summary statistics for the treatment and control groups. The imposed data requirements resulted in a sample of 568 treated students and 1,923 non-treated students whose demographics largely mirrored those of the overall virtual school population.

Table 7: Counts by Student Group in Differences-in-Differences Sample, Treatment vs. Non-Treatment

	Students in Treatment	Percent	Students in Non-Treatment	Percent	Total	Percent
All	568	100%	1,923	100%	2,491	100%
Asian	10	1.8%	56	2.9%	66	2.6%
Black	45	7.9%	175	9.1%	220	8.8%
Hispanic	54	9.5%	172	8.9%	226	9.1%
Am. Indian	19	3.3%	48	2.5%	67	2.7%
White	440	77.5%	1,472	76.5%	1,912	76.8%
English Learners	8	1.4%	62	3.2%	70	2.8%
Students with Disab.	95	16.7%	263	13.7%	358	14.4%
Econ. Dis.	149	26.2%	977	50.8%	1,126	45.2%

Treated students were predominantly white (77.5%), with the next largest group being economically disadvantaged students (26.2%). The control group was also predominantly white (76.5%), and was also significantly (50.0%) economically disadvantaged. Students from non-white ethnicities were underrepresented in both groups.

Results

Table 8: Differences-in-Differences Estimates of the Effect of Virtual School Enrollment on Normalized WKCE Scores

	Reading Score	Math Score
Estimated Treatment Effect	-0.07	-0.19***
	(0.054)	(0.054)
Number of Treated Observations	568	566
Number of Untreated Observations	1,923	1,929
Covariance of Residuals and Treatment	4.6e-16	1.7e-17
Covariance of Residuals and Time Trend	-2.9e-16	2.7e-16
Covariance of Residuals and Estimated ATE	1.6e-17	-2.5e-16

The results of the analysis are shown in Table 8 and show that students who transfered from traditional schools to virtual schools experienced a decrease in reading scores of -0.07 standard deviations, but this estimate was not significant at any level. However, these students also experienced a decrease of -0.19 standard deviations in mathematics performance, and this estimate was significant at the 99% level. For the purpose of verifying compliance with the differences-in-differences parallel-trend assumption, estimates of the covariance between the regression residuals and the individual model parameters are also given. The parallel-trend assumption states that the covariance of each model parameter and the model residuals should be zero (more abstractly, it states that conditional on the covariates, the average outcomes for the treatment and control groups would have followed parallel trajectories in the absence of a treatment), and the data show that the models are in compliance with this assumption (Abadie, 2005).

9 Discussion

Implications for Policy

Research on virtual schools nationwide is alarmingly scarce, and the results of this analysis highlight the desperate need for more studies. As enrollment in digital learning institutions continues to grow nationwide, policymakers have been and continue to be left in the dark when it comes to making informed decisions. This study has attempted to shine some light on the performance of students enrolled full-time in virtual schools, and the primary research questions of this study seem to have been addressed fairly thoroughly: Students who enroll in virtual schools during high school have, on the average, lower probabilities of high school graduation, lower likelihoods of post-secondary enrollment, and experience declines in their mathematics achievement. The quasi-experimental nature of the analysis and the use of closely matched and well-balanced data sets should allow for the attribution of some degree of causality.

The results of this study are very troubling when viewed in the context of virtual school expansion in both Wisconsin and throughout the United States. Among all of the statistical models in this study that examined the impact of virtual school enrollment on the probability of high school graduation, post-secondary enrollment, and standardized test scores, the results were resoundingly negative. In fact, the results suggest that overall, students who enroll in a virtual school at any point during high school are 1-17% less likely to graduate than students who never enroll in a virtual school, and 3-13% less likely to enroll in a post-secondary institution; the strongest statistical model, the balanced covariate model, put these estimates at -17% and -10% respectively.

When examining the effects of virtual school enrollment on standardized test scores in the tested grades, students who switched from traditional public schools to virtual schools experienced declines of 0.19 standard deviations in their mathematics WKCE scores when compared to similar students with similar levels of prior mathematics achievement and on comparable performance trajectories. Although the Estimated ATE on reading scores was also negative, it was not significant at any level. These results suggest that Wisconsin's virtual schools have not yet discovered how to effectively teach mathematics in a virtual environment, and the result is that students who switch to virtual schools immediately have their academic progress slowed.

It is also contended that these results are largely generalizable to full-time virtual students in other states. The rationale behind this contention stems from the fact that student demographics in Wisconsin virtual charter schools mirror those in virtual schools throughout the country (see descriptive statistics of U.S. virtual school populations in Molnar et al., 2013.) Therefore, there is no reason to believe that the results of this study would be limited to the Wisconsin context by demographic features of Wisconsin's virtual schools. In addition, many of Wisconsin's virtual schools are contracted by for-profit companies that have campuses in states across the country, including K-12 Inc., Connections Academy LLC, KC Distance Learning Inc., Insight Schools Inc., Apex Learning Inc., and others (Wisconsin Legislative Audit Bureau, 2010). Working under the assumption that each of these chartering companies use similar versions of their own curriculum in each state in which they operate, there is no reason to think that the results of this study were in any way affected by a curriculum effect specific to Wisconsin.

Despite these findings, however, it is not the intention of this study to declare virtual schools a total failure. Virtual schools can provide educational opportunities to students with special needs who otherwise would not have those opportunities, for example students located in remote rural areas with limited course selection, or students with specific types of learning disabilties. Virtual schools can also offer alternative means of education for children who are victims of bullying or illness, and offer educational choices for parents who simply are dissatisfied with their local traditional schools. But if there's one thing that this study makes clear, it's that Wisconsin's virtual schools have severe issues that need to be addressed before policymakers decide to continue expanding the capacity of full-time virtual charter enrollment.

Shortcomings and Directions for Future Research

One major shortcoming of this study is that data on credit attainment for students, an important predictor of high school graduation, was not available. Many of Wisconsin's virtual schools claim that they often receive students who enter their schools with severe credit deficiency, and are thus already on pace to fail to graduate from high school. Some have also claimed that virtual schools are merely used as a dumping ground for the traditional schools - an alternative learning environment in which to isolate problematic students from the rest of the schools' populations. However, there is currently no solid evidence for assessing the extent of the accuracy of these claims, and though there were no credit attainment data available, this study was able to construct data sets that matched students on a rich set of data including educational attainment, behavior, and parental motivation factors.

Directions for future research should aim to uncover students' motivations for spending parts (or in some cases, the duration) of their secondary education in virtual schools. Do students make the switch willingly, or are the anecdotes about virtual schools as dumping grounds true? Is the decision to attend a virtual school driven by the student, by parents, or by administrators? Additionally, more research is needed on the way in which students in full-time virtual schools learn. Is the curriculum in virtual schools the cause of student academic struggles, or is the problem somewhere in the delivery method? Answers to these questions are currently a mystery, and finding them may shed some light on why virtual school students seem to do so poorly compared with their counterparts who remain in traditional schools.

References

[1] Abadie, A. *Semiparametric Difference-in-Differences Estimators,* Review of Economic Studies, 2005.

[2] Austin, P. *Optimal caliper widths for propensity-score matching when estimating differences in means and differences in proportions in observational studies,* Pharmaceutical Statistics, 2010.

[3] Black, D. and Smith, J. *How Robust is the Evidence on the Effects of College Quality? Evidence from Matching,* Journal of Econometrics, 2004.

[4] Caliendo, M. and Kopeinig, S. *Some Practical Guidance for the Implementation of Propensity Score Matching,* Journal of Economic Surveys, 2005.

[5] Cavanaugh, C. *The Effectiveness of Interactive Distance Education Technologies in K-12 Learning: A Meta-Analysis,* International Journal of Educational Telecommunications, 2001.

[6] Center for Research on Education Outcomes (CREDO). *Charter School Performance in Pennsylvania,* Stanford, CA: Center for Research on Education Outcomes. Accessed 8/29/13 from http://credo.stanford.edu/research-reports.html

[7] Dehejia, R. and Wahba, S. ”Causal Effects in Non-Experimental Studies: Re-Evaluating the Evaluation of Training Programs,” in Rajeev Dehejia, *Econometric Methods for Program Evaluation,* Ph.D. Dissertation, Harvard University, 1997, Chapter 1.

[8] Dehejia, R. and Wahba, S. *Propensity Score Matching Methods for Non-Experimental Causal Studies,* Journal of Economic Surveys, 2005.

[9] Imbens, G. *Nonparametric Estimation of Average Treatment Effects Under Exogeneity: A Review,* The Review of Economics and Statistics, 2004.

[10] Interactive Education Systems Design Inc. *Social Skills of Mainstream Students in Full-Time, Online Public Schools,* 2009.

[11] Kozma et al. *The Online Course Experience: Evaluation of the Virtual High School's Third Year of Implementation,* 2010.

[12] Litke, Eric. ”Virtual charter schools in Wisconsin not making the grade.” *Green Bay Press Gazette* 24 Aug. 2012. Accessed 8/28/13 from http://www.greenbaypressgazette.com/article/20120826/GPG019802/308260104/Virtual-charter-schools-Wisconsin-not-making-grade.

[13] Molnar et al. *Virtual Schools in the U.S. 2013: Politics, Performance, Policy, and Research Evidence.* Boulder, CO: National Education Policy Center. Accessed 8/29/13 from http://nepc.colorado.edu/publication/virtual-schools-annual-2013/.

[14] O'Dwyer, L. and Carey, R. and Kleiman, G. *A Study of the Effectiveness of the Louisiana Algebra I Online Course,* Journal of Research on Technology in Education, 2007.

[15] Pattanayak, C., Rubin, D. and Zell, E. *Propensity Score Methods for Creating Covariate Balance in Observational Studies,* Revista Espanola de Cardiologia, 2011.

[16] Rockman et al. *ED PACE Final Report,* 2007.

[17] Rosenbaum, P. and Rubin, D. *The Central Role of the Propensity Score in Observational Studies for Causal Effects,* Biometrika, 1983.

[18] Setzer, J. C. and Lewis, L. *Distance Education Courses for Public Elementary and Secondary School Students: 2002-03 (NCES 2005-010),* National Center for Education Statistics, 2005.

[19] Wisconsin Legislative Audit Bureau. *An Evaluation: Virtual Charter Schools,* 2010.

[20] U.S. Department of Education, Office of Planning, Evaluation, and Policy Development. *Evaluation of Evidence-Based Practices in Online Learning: A Meta-Analysis and Review of Online Learning Studies,* 2010.

[21] Zhao, Y. and Lei, J. and Yan, B. and Tan, S. *What Makes the Difference? A Practical Analysis of Research on the Effectiveness of Distance Education,* Teachers College Record, 2005.

10 Appendix A

Specification of the Propensity Score Used in Graduation and Post-Secondary Enrollment Analyses

The propensity score was derived from a logistic regression specified as such:

$$Y = \alpha + \beta X + \varepsilon$$

Where βX is a vector of student-level variables *in the year prior to treatment*, including:

MathScore = Student's most recently observed pre-treatment math WKCE score. For students in 11th or 12th grade in the year of treatment, this would be their 10th grade score. For students in 9th or 10th grade in the year of treatment, this would be their 8th grade score.
ReadScore = Student's most recently observed pre-treatment reading WKCE score. For students in 11th or 12th grade in the year of treatment, this would be their 10th grade score. For students in 9th or 10th grade in the year of treatment, this would be their 8th grade score.
AttRate = The student's attendance rate.
Incidents = The number of disciplinary incidents perptrated by the student.
DaysRemoved = The number of days the student was removed from school due to disciplinary incidents.
PercentFRL = The percent of students in the student's school that are eligible for free and reduced-price lunch.
ViolentIncidents = The number of violent incidents occuring in the student's school per 1,000 attendance days.
EconDis = A dummy indicating student is economically disadvantaged.
Disab = A dummy indicating student is a student with a disability.
ELL = A dummy indicating student is an English Language learner.
White = A dummy indicating whether or not the student was White.

The propensity score also includes square and cubic terms of Math and Reading scores and attendance rates, as well as interactions between disciplinary incidents, days removed, and attendance rates with the White dummy variable.

11 Appendix B.1

Table 9: Student Counts and Percents by Subgroup, Matched Data Sample for Estimation through Balanced Data

	Number of Students: Treatment	Percent	Number of Students: Non-Treatment	Percent
All	1,516	100.0%	1,516	100.0%
Asian	21	1.4%	16	1.1%
Black	78	5.1%	74	4.9%
Hispanic	64	4.2%	93	6.1%
Am. Indian	38	2.5%	16	1.1%
White	1,315	86.7%	1,317	86.9%
English Learners	8	0.5%	53	3.5%
Students with Disab.	189	12.5%	167	11.0%
Econ. Dis.	552	36.4%	617	40.7%

Table 10: Covariate Balance of Aggregate Sample, Regression Method

	Mean Treatment	Mean Non-Treatment	P-Value
Reading Score	-0.04	-0.10	0.06
Math Score	-0.21	-0.27	0.04
Attendance Rate	83.19%	84.32%	0.05
Disciplinary Incidents	0.36	0.43	0.06
Days Removed from School	2.31	1.35	0.03
Percent White	0.87	0.87	0.91
Percent Econ	0.36	0.41	0.02
Percent Disab	0.12	0.11	0.21

12 Appendix B.2

13 Appendix C

Sample Means and Statistical Test of Difference in Means by Strata of Sample Derived for the Balanced Covariate Estimation Method

Table 11: Student Counts and Percents by Subgroup, Matched Data Sample for Estimation through Balanced Data

	Number of Students: Treatment	Percent	Number of Students: Non-Treatment	Percent
All	1,516	100.0%	1,516	100.0%
Asian	21	1.4%	16	1.1%
Black	78	5.1%	74	4.9%
Hispanic	64	4.2%	93	6.1%
Am. Indian	38	2.5%	16	1.1%
White	1,315	86.7%	1,317	86.9%
English Learners	8	0.5%	53	3.5%
Students with Disab.	189	12.5%	167	11.0%
Econ. Dis.	552	36.4%	617	40.7%

Table 12: Covariate Balance of Aggregate Sample, IWP Method

	Mean Treatment	Mean Non-Treatment	P-Value
Reading Score	-0.04	-0.10	0.06
Math Score	-0.21	-0.27	0.04
Attendance Rate	83.19%	84.32%	0.05
Disciplinary Incidents	0.36	0.43	0.06
Days Removed from School	2.31	1.35	0.03
Percent White	0.87	0.87	0.91
Percent Econ	0.36	0.41	0.02
Percent Disab	0.12	0.11	0.21

Table 13: Covariate Balance Stratified by Propensity Score, Full Matched Data Sample

	Mean Treatment	Mean Non-Treatment	Strata	QUANITLE	Sig
Reading Score	0.13	-0.03	0.06	1	
Math Score	0.11	0.00	0.14	1	
Attendance Rate	96.91%	96.59%	0.38	1	
Disciplinary Incidents	0.14	0.08	0.31	1	
Days Removed from School	0.21	0.50	0.44	1	
Percent White	0.86	0.78	0.02	1	*
Percent Econ	0.26	0.30	0.41	1	
Percent Disab	0.14	0.10	0.15	1	
Reading Score1	0.04	-0.01	0.51	2	
Math Score1	-0.13	-0.16	0.74	2	
Attendance Rate1	94.92%	93.93%	0.10	2	
Disciplinary Incidents1	0.11	0.09	0.66	2	
Days Removed from School1	0.41	0.13	0.25	2	
Percent White1	0.89	0.89	0.88	2	
Percent Econ1	0.27	0.29	0.61	2	
Percent Disab1	0.10	0.12	0.46	2	
Reading Score2	-0.06	-0.03	0.61	3	
Math Score2	-0.27	-0.26	0.86	3	
Attendance Rate2	90.86%	90.64%	0.60	3	
Disciplinary Incidents2	0.25	0.21	0.52	3	
Days Removed from School2	0.39	0.48	0.50	3	
Percent White2	0.86	0.87	0.89	3	
Percent Econ2	0.40	0.45	0.26	3	
Percent Disab2	0.15	0.10	0.09	3	
Reading Score3	-0.02	-0.21	0.03	4	*
Math Score3	-0.28	-0.41	0.10	4	
Attendance Rate3	85.49%	85.41%	0.87	4	
Disciplinary Incidents3	0.44	0.41	0.66	4	
Days Removed from School3	1.08	0.95	0.59	4	
Percent White3	0.92	0.86	0.05	4	
Percent Econ3	0.40	0.50	0.03	4	*
Percent Disab3	0.13	0.15	0.50	4	
Reading Score4	-0.19	-0.16	0.78	5	
Math Score4	-0.28	-0.36	0.28	5	
Attendance Rate4	67.43%	70.43%	0.03	5	*
Disciplinary Incidents4	0.60	0.73	0.35	5	
Days Removed from School4	4.68	2.31	0.08	5	
Percent White4	0.89	0.92	0.16	5	
Percent Econ4	0.41	0.42	0.85	5	
Percent Disab4	0.12	0.14	0.49	5	