Drivers of Student Satisfaction and Student Outcomes in K-12 Online Learning Environments

Jamie John Beaton

Blavatnik School of Government

University of Oxford

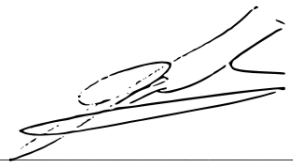
A thesis submitted for the degree of

Doctor in Philosophy

2021

Signed Declaration

I, Jamie J. Beaton, confirm that the work presented in my thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis.

Signed: 

Date: 19th September 2021

Acknowledgements

I am very grateful to the group of people that have supported me in the completion of this thesis and the participation of many students, teachers and policy makers who offered their time to play a part in the studies presented.

A major thanks to Professor Lucy Bowes who advised my thesis. Her deep knowledge of psychology, education and experimental design was critical to many aspects of this thesis. Lucy, thank you for your endless faith, support and encouragement of my work and belief in me to embark on this exciting academic challenge. Thank you for keeping me on track, challenging me to learn new techniques and approaches and ultimately producing an exciting mixed methods thesis that has made me a more critical thinker. Your support vastly exceeded any expectation I ever had coming in of a DPhil advisor and I have always looked forward to all our time spent together and benefited immeasurably from your insights, advice and guidance.

A big thank you to Professor Peter Kemp who played a critical role in advising me on the public policy aspects of this thesis so that all of my hard work could be packaged in a useful guide for regulators and lawmakers as they seek to understand the emerging online schooling space. Thank you to Professor Pepper Culpepper for guiding me into my journey as a DPhil student and exposing me to many useful academic resources, techniques and insights that helped me navigate this chapter. Thank you to Professor Sue Dopson, Professor Julienne Labonne and Professor Jane Gingrich for your brilliant feedback along the way.

Thank you to Mary Eaton for her warmth, support, check-ins and inspirational stories throughout my journey at Oxford. I always felt at home in your office sipping tea and sharing the progress I was making. You are the heart and soul of Rhodes House and you make our community stronger.

Thank you to the New Zealand Rhodes Selection committee for having faith in me to select me to go on this amazing academic journey that has challenged me in new and fruitful ways. I consider your faith in me as the highest honor of my life and will be forever grateful that you chose to bet on me.

Thank you to Julian Robertson for funding the New Zealand Rhodes Scholarship which enabled me to complete these academic studies. Thank you also to Julian for mentoring me since the age of eighteen and helping me to learn about the economy,

politics, education, government and many aspects of life that have made me a stronger person. Your impact on many facets of my life will always be cherished.

Thank you to Jhett Koo for your brilliant insights, feedback and support on this research, in particular, in helping to facilitate the randomized control trial within Crimson.

Thank you to Dr. Galen Buckwalter. What started as a chance encounter at 101 Park Avenue turned into a brilliant friendship, a shared fascination with the role of personality in education and years of exciting research. I admire your independent thinking and novel contributions to the matching literature both academically and throughout industry.

Thank you to Tomohiro Hoshi and the Stanford Online High School. Under your leadership, you have created a major innovation in learning and it was fascinating to understand more about the unique environment you have created and its potential to impact more students in the future.

Thank you to John Morris, Fraser White, David Mansfield, Mark Phillips, Karen White, Karan Khemka, Larry Summers, Nikki Kaye, Bror Saxberg and the many other individuals who helped me in various ways in my journey to learn about the online schooling (and schooling) space in general.

Thank you to my partner Nensi for cheering me on as I passed through each milestone and learning on this journey.

Finally, thank you to my mother Paula, father Glenn, granddad John and grandmother Sarah for believing in me since I popped into the world and providing me the emotional support necessary for me to keep pushing through this work. My love for learning comes from you Mum.

Abstract

My thesis examines what drives student outcomes and student satisfaction in online schooling. Specifically, I examine whether psychometric matching between students and tutors can be used to enhance outcomes or satisfaction and in light of this, the key considerations for governments seeking to regulate the online schooling landscape.

Two systematic literature reviews are conducted: firstly, what drives student outcomes and student satisfaction in online schooling and secondly (Chapter Two), the use of algorithmic matching between students and educators in schooling to improve student outcomes and/or student satisfaction (Chapter Three). I perform a qualitative analysis of leading virtual high schools, interviewing 21 students using NVivo qualitative analysis software (Chapter Four). I conduct cross-sectional analysis on a data-set from Crimson Education, an online education company, to evaluate the impact of psychometric characteristics on student satisfaction (Chapter Five). I run a randomized control trial to test the causality of a psychometric matching algorithm on student outcomes and student satisfaction (Chapter Six). My randomized control trial analysis finds that psychometric matching can lead to significant impact on writing scores and on student satisfaction within Western Europe and Latin America. My cross-sectional analysis also finds that student satisfaction scores can be impacted by the psychometric characteristics of students. Finally, I provide a set of recommendations for public policy makers looking to adapt legislation to the emerging online high schooling sector in light of my findings from earlier chapters, in particular my qualitative analysis and systematic reviews (Chapter Seven).

Contents

1 Introduction

1.1 Introduction

As a graduate student at Stanford studying a Masters in Education, I came across the Stanford Online High School (SOHS). SOHS is ranked as one of the top 10 high schools in America but is fully online (Niche, 2020). I was curious how a fully online school could generate such a level of academic attainment and also about how widespread online schooling is. I began to research the space and soon came across chains like K-12 Inc (Stock Ticker: LRN) which generates more than a billion USD in revenue providing virtual schooling solutions to families across America. All in all, approximately 1% of American high school students attend an online high school (National Centre for Education Statistics, 2020). Interestingly, despite the growth in virtual schools, a tertiary to K12 lag is occurring where adoption of online learning is dramatically higher at the university level. According to the Centre for Online Education, in 2012, 46% of four year colleges offered online programs, but by 2014 that had jumped to nearly 60%. Virtual schooling is of particular interest given the relative nascence of academic literature supporting the model but its rapid growth in markets such as the United States. In Keeping Pace With Online Learning (2014), it was reported that enrollment grew 6.2% from 2013 to 2014 with approximately 315,000 students in the US enrolled in entirely online high

school programs. Online K-12 Education is one of the fastest growing education reforms in America at present (Watson et al, 2014).

According to Tomohiro Hoshi, Head of Stanford Online High School, approximately one in six Americans complete their degree online. This implies that online schooling has a more than 15 times higher penetration rate in Tertiary than K12. The sustained growth in online schooling at the K12 level could continue to grow dramatically for decades to come given the demonstrable ability for online to become a major delivery channel, even at much higher price points, in the tertiary sector. I began this thesis before COVID-19 but the evolution of the pandemic has thrust debates about online schooling into the mainstream.

Virtual schooling growth has disproportionately occured in the United States, but many other OECD countries and developing nations do not have public policy frameworks in place to support virtual schools with neither private-pay or government funded options.

Furthermore, technology development is only likely to accelerate this adoption. Virtual schooling in the USA began at a time where video calling was in its infancy, Facebook hadn't launched yet and virtual reality and augmented reality existed only in science fiction movies. Today, platforms like Zoom, a video-conferencing solution, have a market capitalization of $44 billion USD (May, 2020) and ~300 million users. Video conferencing, a necessary technology to support successful synchronous online schooling, has a wide variety of vendors from Skype, Cisco, Google Hangouts, Zoom with new entrants including companies like Facebook. Remote work platforms like Slack offer the ability for entire companies to collaborate virtually. Features such as virtual backgrounds help students to showcase their personality in a similar way to eccentric clothing choices or distinctive haircuts. The option to record sessions provides virtual schools with enhanced child-safety opportunities, students with the ability to re-play confusing components of class and a broad incentive for all participants to be on their best behavior. Cloud computing platforms such as Amazon Web Services provide transcription services that enable class lessons to be easily searched by students to find specific mentions of a concept. Facial recognition

technology enables sentiment analysis to suggest the mood of various students and to guide a teacher to be able to determine if a student is paying attention, disengaged, sad or happy among other things. Virtual reality provides opportunities for remote practical science education. Social media networks enable the creation of a sense of belonging without necessarily physical proximity between learners and teachers. As with any technology-driven innovation, the various benefits may come alongside a variety of costs.

The niche virtual schooling market, which was serving less than 1% of Americans prior to commencement of this thesis, has become a live experiment for entire nations forced to rapidly find alternative delivery models for their students as COVID-19 has become a global pandemic. Physical schools are unequipped with a robust technology platform, do not have dedicated e-learning resources, have teachers that are not trained for the unique challenges that online teaching requires and do not have the necessary tracking in place to measure attendance, engagement or progressive academic performance. It is clear that schools going forward will need to be equipped to deliver virtually or may need to acquire competencies in these areas. Additionally, it is likely that existing and new virtual school operators are likely to grow substantially as parents have now had exposure to this delivery model. Recent entrances to this market include Harrow School Online, Valenture Institute, Dulwich International (a major operator of private schools in China) and Inspired (a major global operator) are actively looking at expansions into these areas.

Despite the rapid growth in virtual schooling before COVID-19 and the unprecedented use of online learning during COVID-19, the National Education Policy Centre remarks "More than twenty years after the first virtual schools began, there continues to be an inadequate research base of empirical studies to guide the practice and policy of virtual schooling,". My thesis aims to contribute to this deficiency in the academic literature and provide a catalyst for future analysis of the online learning sector that is evidence-based and can help guide regulators in their evaluation of this sector and school operators in creating high-performing learning environments that enhance both student outcomes and student satisfaction. My

thesis also aims to offer specific recommendations around the use of psychometric matching to enhance both student outcomes and student satisfactions, which is particularly applicable in virtual schooling but can also be applied more generally.

While I am particularly focused on virtual schooling, UNESCO estimates that around 1.29 billion people are currently enrolled in primary and secondary schools around the world. According to the World Bank, the average country spends around 4.88% of GDP on education. With such massive levels of investment going into schooling, improvements in student outcomes and student satisfaction are of significant importance. According to Dr Bror Saxberg, Chief Learning Officer of the Chan Zuckerberg Initiative, virtually all students are paired with teachers on the basis of availability or schedule, as opposed to any personalized algorithmic matching procedure (Saxberg, 2019). In small schools, administrators will often have limited ability to choose which teachers are assigned to which students based on human resource constraints. In larger schools, where there exists a number of teachers who could potentially teach a given class, allocation based on schedule alone does not necessarily maximize learning outcomes. Like many other domains which have successfully implemented algorithmic matching such as online dating, I hypothesize that by considering the individual characteristics of students and teachers, class composition can be more effectively optimized and translate to higher quality satisfaction and outcomes.

My systematic review finds that key drivers of student outcomes and student satisfaction in virtual schooling are (1) parental engagement, (2) self-motivation in students, (3) strong relationships with mentors and (4) peer effects. The systematic review also suggests that synchronous (live teacher) delivery models, in general, may be more beneficial for students than the asynchronous, video-lecture based models that defined early endeavors in this space. Key research gaps that need to be explored include the differential impact of class size in an online environment compared to a physical environment, the impact of "tracking" the online classroom experience on learner and teacher effectiveness, the differential impact of differing levels of parent engagement on student outcomes and the impact of age on effectiveness

of virtual schooling. The research in this space is relatively nascent and generally only features US studies to date.

I aim to catalyze a wave of research into how schools can more effectively leverage their teaching resources by systematically pairing students and teachers together. Our specific analysis will focus on two key variables for algorithmic matching: personality characteristics, as determined by a personality assessment known as "HEXACO," and gender.

This DPhil is grounded in the discipline of Education in the context of K-12 Virtual Schooling. The specific sub-fields I address are the (1) education matching literature, which presently pertains to ethnic-racial and gender matching as well as (2) drivers of student outcomes and student satisfaction in online learning. My work in the first sub-field extends the burgeoning literature, which employs the use of psychometrics to systematically identify, match and optimize mentor-mentee interactions in the context of education. The education matching literature is a relative of the broader matching literature. Applications of matching have included allocation of students to public schools in the Boston region (Kominers et al, 2016). The matching literature broadly attempts to design efficient mechanisms to allocate scarce resources, often with the use of various algorithms including deferred acceptance but traditionally the work of matching to boost outcomes inside a marketplace has been primarily conducted by enterprises. My work in the second sub-field offers individuals trying to get up to speed on innovations in online schooling a primer on developments in the space, qualitative analysis of the drivers of outcomes and satisfaction within one of the original online high schools (Stanford Online High School) and a set of related public policy recommendations regulators can use in navigating this evolving space.

By leveraging data from Crimson Education, I will be able to provide a much-needed large sample size analysis to advance the matching literature, specifically in the online learning domain. Much of the existing literature employs case studies and qualitative analysis of factors, which may lead to a promising student-teacher interaction. Additionally, most of the literature does not collect a high frequency

of student satisfaction feedback for a given student-teacher pair and lacks a clean measure of student outcomes. Finally, it is rare to find data which addresses the case of one-on-one matches at a large scale. My research will provide virtual schools, which typically cater to a large population size such as K12 Inc, Stanford Online High School as well as large physical school chains, valuable insights into how they can better serve their students with existing resources. This research will also be relevant to private tutoring companies and mentoring organizations that attempt to match individuals with students to improve their achievement, engagement or satisfaction in the education system. Both physical and online schooling rely heavily on one-on-one student-teacher relationships both formally and informally to enhance the experience. For online high schools, one-on-one interaction through office hour sessions is a particularly important part of the learning experience given the limited opportunity for informal interaction outside of class. My analysis, which focuses on optimizing one-on-one matches, will be directly relevant to these interactions and will also likely yield insights that can be applied to larger groups (Bettinger, 2018).

As discussed earlier, many governments in the OECD including New Zealand, a region of focus in this thesis, are proactively looking at regulation to facilitate online learning. Depending on the effectiveness of algorithmic matching and the marginal ability to deploy such algorithms to online high schools compared to physical high schools, it could change the public policy appetite to support the fledgling virtual high-schooling market. Additionally, many Ministries of Education are experiencing rising education costs, primarily from labor, without commensurate improvements in student outcomes. A systematic matching process, that can be recommended to school bodies, could help to enhance the government's return on education outcomes for every marginal dollar invested in the schooling system. I aim for this research ultimately to play a meaningful role in opening up policy makers' perspectives to broader personalized learning innovations, which leverage a more granular understanding of the student and the teacher to outperform traditional classroom environments.

1. Introduction

I opted for a mixed-methods research approach to combine the rigour of quantitative methodologies like randomized controlled trials with the critical background context, user empathy and more subtle relationships of qualitative analysis. I decided on this approach given the relatively novel virtual schooling space. Many traditional researchers may only be familiar with offline contexts so a deep qualitative section that helps paint why users opt into this type of schooling may help people transition into a stronger intuitive understanding of online schools. I began with a systematic review of what drives student outcomes and student satisfaction in virtual schooling to provide broad coverage of the various factors that are necessary to understand the state of the literature so far. I then conducted a systematic review of education matching because the field is relatively underdeveloped given the extensive academic progress in other domains of the matching literature. This was particularly intriguing given the larger availability of data from virtual learning environments (given data from the live class interactions) making a richer set of matching algorithms possible. After conducting this systematic review, it became clear that both gender and ethnicity had been addressed in the literature to some extent but cognitive and psychometric characterics had not. This was an area that was challenging to explore given the lack of availability of psychometric data in schools. By obtaining a unique dataset from Crimson Education, I was then able to conduct various quantitative analyses of the impact of psychometric matching on student outcomes and student satisfaction. These insights, while impactful and a meaningful contribution to the literature, are hard to reconcile with the broader virtual schooling landscape in isolation. To provide context, particularly to readers new to the virtual schooling landscape a qualitative study was conducted of students from leading online schools including the Stanford Online High School. Finally, public policy recommendations were made, primarily drawing from the systematic review of student outcomes and student satisfaction as well as the qualitative study.

1.2 Research Question

The existing literature presents a compelling opportunity. Much of the existing analysis of drivers of student outcomes and student satisfactions has been done in the context of traditional brick-and-mortar school environments. Many aspects of research into traditional schools can be directly applied to virtual schools. Ultimately, educators are optimizing for the same outcomes in students. Students are still interacting with one another albeit in digital environments. Students are learning the same curriculum and preparing for the same examinations. As a result, the primary differences occur when one considers the varying characteristics of students who perform well in what environment. Do students perform equally in both? Do students with certain characteristics perform better in online schooling and other students perform better in brick-and-mortar schooling? Is brick-and-mortar schooling, virtual schooling or "blended" schooling in which students engage in both styles of delivery more optimal on average?. Virtual schooling tends to provide opportunities to escape some of the traditional constraints of physical schools around teacher availability, cost structure of facilities and other non-teacher expenses, and access to more diverse communities. The central debate about the role of virtual schooling comes down to whether any trade-offs in the effectiveness of the live-class learning environment online versus physical and access to physical social interactions with school peers is worth it when considering some of the practical benefits of virtual schooling.

In the context of matching, race-based matching has been analyzed in the context of supporting achievement of underprivileged minorities but the general applicability of this type of matching for other groups such as European or Asian students has not been well verified. Gender-based matching has been considered with plenty of ambiguous evidence around age groups, mechanisms underpinning observed outcomes and statistical significance of findings. Psychometric-based matching has had fairly limited coverage in the literature but the direct analysis so far appears promising. This thesis is being conducted in a time defined by

rapidly expanding demand for online high schooling, despite academic evidence remaining broadly inconclusive about its efficacy. Finally, much of the existing analysis is based on case-studies or observational data, which have not been able to use large sample sizes to generate robust empirical findings. Based on this and the motivations mentioned, my research question is as follows:

What drives student outcomes and student satisfaction within online schools? Specifically, can systematic algorithmic matching between students and teachers using psychometric characteristics and/or gender improve student outcomes and student satisfaction in an online learning environment?

The question seeks to test the current lack of systematic matching, whereby students are matched based on schedule and availability with no consideration of a student's personality, background or learning outcomes. The question focuses on one-on-one interactions, which simplifies the matching algorithms required by ignoring group dynamics and peer-to-peer learning interactions. This is designed to proxy the student-teacher mentoring relationship, which often exists within schools and has direct relevance to instances such as office hours and tutoring.

1.3 Methodology

My thesis utilizes a mixed-method approach to provide insight into the key drivers of student outcomes and student satisfaction in online schooling and evaluates one potential driver, algorithmic matching of gender and psychometrics, to enhance student outcomes and student satisfaction. Finally, I share public policy implications and recommendations from the research.

My thesis is structured as follows:

Chapter One: Abstract/Introduction

A holistic overview of the thesis, key research questions, context behind the research question and methodology.

Chapter Two: Systematic Review - Online Schooling Satisfaction & Outcomes

A summary of the findings of existing literature which tests which factors drive student outcomes and student satisfaction in online schooling.

Chapter Three: Systematic Review - Algorithmic Matching

A summary of the findings of existing literature which tests various ways in which students and teachers have been matched (gender, ethnicity, cognitive-style) and the impact of these techniques on outcomes.

Chapter Four: Qualitative Analysis of Drivers of Online Schooling Utilization, Outcomes and Satisfaction

The findings of an extensive range of interviews with learners from the Stanford Online High School and other online high schools are presented, qualitative analysis is conducted and key themes are presented.

Chapter Five: Cross-Sectional Analysis

A cross-sectional study is conducted to evaluate the impact of student and tutor psychometric characteristics on student satisfaction scores in an online learning environment. Linear regression models are developed in one sample and tested against an independent sample to test the impact of various psychometric characteristics.

Chapter Six: Randomized Control Trial

A gold standard randomized control trial is conducted to evaluate the impact of psychometric matching between students and tutors on student outcomes (measured by SAT score improvement) and student satisfaction (measured by session review scores). The impact of gender matching between students and tutors is also considered.

Chapter Seven: Public Policy Recommendations

Findings from my thesis, in particular, from the qualitative analysis (Chapter Four) and systematic review of student outcomes and satisfaction (Chapter Two) are used to inform a set of public policy recommendations for education regulators.

Chapter Eight: Conclusion

The key findings from my mixed-method thesis are summarized.

My mixed-method approach is a novel contribution to academic literature in this field. Firstly, there have been no randomized control trials to date in the psychometric matching literature and very few of any kind in the broader education matching literature. This is partially because of tendencies in the field towards qualitative research but also because it is rare for any school groups to capture any kind of session by session feedback or be able to deploy broad psychometric testing in a capacity. Finally, it is very rare to have a meaningful like-for-like outcome measure like SAT testing deployed and assessed pre and post an intervention in a systematic way.

My randomized control trial deploys a psychometric matching algorithm (from the assessment HEXACO) to match students and tutors by personality in the intervention group compared to a control group that is matched without consideration of personality. The process of running the randomized control trial was arduous and required extensive experimental design work, nuanced implementation considerations to ensure anonymity of students and then a large sample size. In the education matching literature, there was previously only qualitative commentary of the use of personality matching so a large-scale randomized control trial is a major step forward. I chose to use a randomized control trial as it is regarded as the gold standard of experimental design and is able to measure causal relationships which I hypothesized existed between psychometric matching, student outcomes and student satisfaction.

I was able to conduct a large cross-sectional analysis on historical data evaluating students and tutors that had conducted the HEXACO psychometric assessment against student satisfaction scores. This analysis yielded a variety of statistically significant findings on the impact of student HEXACO scores on student satisfaction as well as the impact of various psychometric compatibility algorithms comparing the difference in personality traits between students and tutors. I was then able to test these models against the new sample generated from the randomized control trial to check the robustness of the findings and found a number of models, particularly

relating to student HEXACO scores, that continued to be statistically significant out-of-sample. Psychometric assessments involve a significant number of independent variables and formulation of robust models from existing research into specific traits alone is challenging. Data-driven correlational analysis has clear weaknesses in being unable to prove causality. I strengthened my methodology by repeating my analyses in a second, independent data set to increase confidence in the findings in terms of replicability.

Thirdly, I built a strong relationship with the Stanford Online High School (SOHS), enabling several days of on-site interaction with Tomohiro Hoshi, the Head of Stanford Online High School and qualitative interviews with twenty-one virtual schooling students. SOHS is the world's leading virtual school by student outcomes ranked in the top 10 in the US of any school (Niche, 2020). These interviews were then analyzed using NVivo software to more rigorously understand the key words, variables and relationships that were commonalities across the discussions. It was important to have this qualitative analysis because it incorporates the lived experience of online schooling students and gives me and others reading this work a grounded reality into the research being conducted and who it affects.

Fourthly, I conducted two systematic reviews. No previous systematic review has been conducted in the education matching literature and very few have been conducted in the literature on online schooling outcomes and satisfaction. While systematic reviews are more common in fields such as medicine and social sciences, I chose this technique because it enables a comprehensive, back-testable, audit of the entire literature indexed to key words and avoids the bias that general literature reviews have. A general literature review allows the researcher to selectively choose pieces that warrant exploration but this allows for bias in article selection. In a sector like education where educators hold varied opinions, a qualitative literature review seemed suboptimal in providing a balanced analysis of the literature for the benefit of policymakers and educators.

Fifthly, I provided various public policy recommendations derived from our systematic reviews, qualitative analysis, cross-sectional analysis and randomized control trial supported by case studies of the for-profit online university sector.

1.4 Primary Findings

The systematic review of student outcomes and student satisfaction in virtual schools found key drivers to be (1) the impact of parental involvement, (2) the importance of self-motivation, (3) the impact of peer effects, (4) the impact on general student achievement and (5) the importance of mentorship in online learning environments.

The systematic review of education matching showed (1) a lack of impact of matching by gender and (2) a lack of impact of matching by ethnicity but broadly a sparse literature that needed more contribution. I also noted a sparsity of analysis of personality-based matching.

Our qualitative analysis found the following recurring themes in students' perceived requirements of success of virtual learning: (1) self-motivation, (2) desire for academic acceleration through extension coursework, (3) less stressful social and competitive environment (4) flexibility in schedule facilitating other activities. The core themes we found in students' perceived weaknesses of online schooling were: (1) higher barriers for communication with teachers, (2) less spontaneous opportunities for social interaction, (3) limited access to school extracurriculars.

The randomized control trial produced a number of important findings. The key findings were that (1) matching by the personality dimensions algorithm lead to a statistically significant increase in student session ratings in Western Europe and Latin America, (2) matching by the personality dimensions algorithm was associated with a statistically significant decrease in student session ratings in Asia and Eastern Europe, (3) matching by the personality dimensions algorithm lead to a statistically significant improve in SAT writing performance but not on reading or mathematics, (4) gender matching lead to a statistically significant increase in average session rating and (5) gender matching lead to a highly statistically

significant impact on the first two lessons of a student-tutor pair and then waned in significance as the number of interactions grew.

Other key findings were that (6) student outcomes measured by centred SAT score improvement was not correlated with student satisfaction scores measured by average session rating, (7) tutor personality traits had a minimal first order impact on average student session ratings, (8) student personality traits agreeableness and openness to experience are positively correlated with increased student ratings and these findings are robust across two independent data-sets, (9) session ratings between students and tutors are positively correlated with HEXACO personality sub-dimension Sincerity when the ratings are weighted based on meeting frequency and (10) the relative difference between the HEXACO personality sub-dimension Inquisitiveness is positively correlated with average session score.

In my public policy analysis, I suggested (1) broader adoption of evidence-based outcome measures such as randomized control trials should be used by public policy makers as the status quo, especially in countries like New Zealand. (2) I also suggested public policy makers should proactively develop legislation to support the emerging trend of online high schooling particularly post COVID-19 and gave various suggestions for consideration.

Altogether, these findings show that systematic matching of students and teachers online is a worthwhile area for future research and experimentation. Systematic personality matching based on a consideration of the traits of the student and the tutor may improve or worsen student outcomes depending on the algorithm in question. Gender matching appears to play a role in student satisfaction, especially early in the student-teacher relationship formation. More broadly, online schooling appears to be an attractive option for a specific cohort of learners. This cohort is likely to continue to grow in a post COVID-19 era and it is worthwhile for public policy makers to proactively consider the needs of this group in the design of their regulatory environment for K-12 schooling going forward.

2

Systematic Literature Review: Students Outcomes and Student Satisfaction in Virtual Schools

2.1 Introduction

Traditionally, literature reviews in the fields of public policy research were written as "expert reviews". Non-systematic reviews have been shown to increase the risk of bias (Ferrari et Al, 2015) around article selection and interpretation. A major focus of this thesis is to be balanced, data-driven and objective in assessing evidence around online education. As a result, I have chosen to conduct a systematic review. A systematic review, typical of the medical sciences for applications such as drug efficacy research, always involves pre-specification of relevant search terms around a tightly scoped set of questions to locate relevant articles. The review then covers all articles that have been published that meet this search criteria and it usually ranks them using a quality score index in order to weight the applicability of their findings. While there is some variance in the implementation of the systematic review, the key priority is to be objective about the articles produced. Systematic reviews are methodical and repeatable. The term "systematic" most clearly refers to the idea that by following a set of pre-specified procedures, future researchers

could replicate the findings in their entirety. The most useful review synthesizes studies to extract broad theoretical conclusions about what a given literature means, linking theory to evidence and evidence to theory. This methodology is becoming increasingly common in the social sciences (Kang et al, 2019).

While the mechanical systematic review has a number of strengths it is not without its limitations. Firstly, a systematic review may be susceptible to missing certain important papers based on subtle changes in key language used to describe the paper if these do not appear in the search terms chosen. Secondly, important papers that may be in other languages or from other disciplines with subtle but important relationships to the research question may be excluded. Thirdly, the general focus on quantitative studies in systematic reviews given the heritage in medical research may unnecessarily penalize important findings that are more qualitative in nature. This was particularly true in this analysis because some fairly useful papers for my analysis received relatively low scores in the objective quality score assessment. In order to address these, it is useful to conduct a systematic review and then interweave the findings with more associative approaches. This is most appropriate in the discussion section so the analytical precision of the systematic review method is not compromised.

My systematic review focuses heavily on student outcomes and student satisfaction in the context of virtual schools but it is important to address the large literature on efficacy in physical schools in general. Much of this literature is relevant to online schooling. While this is outside the scope of this systematic review, notable contributions include Professor John Hattie's effect size literature in which he adjusts and ranks a variety of drivers of efficacy in an illuminating meta-analysis. He finds collective teacher efficacy, self efficacy, teacher credibility, teacher estimates of achievement, micro-teaching and video lesson reviews as examples of some of the most effective drivers of academic performance (Hattie, 2018). Bandura was another notable contributor who analyzed teacher efficacy by assessing teachers across various dimensions like self-organization, goal selection, self-reflection, resilience and stress management and found strong relationships between highly efficacious teachers and

student outcomes (Bandura, 1977). Student satisfaction has been well analyzed in "Towards a Comprehensive Student Satisfaction Model" and outlined perceived educational service quality and value for money as two key drivers when a student is embarking on a paid program such as a private school education (Arora et al, 2021).

The objective of this systematic review is to analyze what factors could have a correlation with student outcomes and student satisfaction in virtual schools. In reviewing correlation analysis, I cannot make causal claims but I can synthesize the evidence to identify consistent patterns of research findings and estimate the strength of relationships across multiple studies. This systematic review provides new researchers to the virtual schooling space a robust introduction to the key drivers of student outcomes and student satisfaction. It also helps provide the context that the findings from qualitative study later in the thesis (Chapter Four) are relatively aligned with existing work.

Most education researchers have focused on traditional, brick-and-mortar schools. There have been some systematic reviews of online schooling such as "Research and Practice in K-12 Online Learning: A Review of Open Access Literature" (Cavanaugh, Clark et al, 2009) but there are important gaps. For example, there has been limited work conducted on the use of matching algorithms between students and teachers to improve student outcomes. Much of the large scale randomized control trials have occurred in physical schools around the world and a significant proportion of the world (pre-COVID) learned in fully offline brick-and-mortar schools. In performing my research in online learning, a useful starting block is to understand what appears to be working at the moment. It is also useful to understand what types of academic outcome variables are measured, how student satisfaction is measured and to gain familiarity with the types of experiments that have been run before. These insights guided the experimental design of the randomized control trial as well as the public policy recommendations. I aim to make this systematic review thorough yet succinct, so a researcher looking to quickly make the transition into online education experiments can use this as an introductory primer to the work done so far.

This chapter will first outline (1) the question addressed in the systematic review, (2) the methodology used in the systematic review, (3) the summary tables of the systematic review, (4) key findings of research papers assessed and (5) discussion of the key themes emerging from the review.

For the purposes of this research, I define K-12 as learning that takes place between kindergarten and 12th grade in the US system, and schooling as formal education activities designed to enhance student outcomes that occur on a regular basis.

2.2 Systematic Review Logic

A systematic review is a rigorous compilation of evidence from all primary research studies within a defined set. I have chosen to do a systematic review given the scarcity of the literature, the reduced human bias in the process, increased repeatability because of the explicit methodology and a systematic presentation of synthesis of the characteristics and findings of the included studies.

Across the set of literature, there are several case control studies in schools which involve large sample sizes that are fairly rigorous, but there is an absence of literature that holistically ties the themes together from this existing work. Some existing literature reviews include "The Effects of Distance Learning on K-12 Student Outcomes: A Meta-Analysis (Gillan, Kromney et al, 2004), "Online Learning: Adoption, Continuance and Learning Outcome, A Review of the Literature" (Sharma et al, 2018) and "Research and Practice in K-12 Online Learning: A Review of Open-Access Literature" (Clark, Barbour et al, 2009). These reviews are all narrative in nature and tend to only address student outcomes narrowly without viewing them alongside measures of student satisfaction in order to draw an overarching conclusion about the mechanisms at play. Additionally, the methodology by which papers were assessed for quality varied considerably throughout the reviews, so it was challenging to ascertain quality of the underlying articles without reading them in-depth.

Given the existing virtual schooling research, a systematic review is useful given the lack of current systematic literature reviews conducted in this space around

student outcomes and student satisfaction in virtual schooling. This review will provide future researchers with a clear survey of the literature that is repeatable and with clear, objective quality indicators. It will also cover both student outcomes and student satisfaction simultaneously in order to tease out a more comprehensive explanation for the factors at play.

The following search terms and inclusion/exclusion criteria will be used for our systematic review. This search was performed in November 2018.

("Student" OR "Students" OR "Pupil" OR Child") AND ("Satisfaction") AND ("Virtual" OR "Online" OR "Supplementary") AND "School"

(("Student" OR "Students" OR "Pupil" OR "Child") AND ("Outcome" OR "Achievement" OR "Progression") AND ("Virtual" OR "Online" OR "Supplementary") AND "School"

2.3 Questions addressed in systematic review:

The systematic review seeks to answer the following questions:

a. Specifically in online schooling, what characteristics have been shown to improve student outcomes?
b. Specifically in online schooling, what characteristics have been shown to improve student satisfaction?
c. What relationships exist between student outcomes and satisfaction?

The search questions specifically focus on online schooling because there are likely to be some techniques that may work well in offline schools for driving student outcomes that have no relevance to online schooling. An example could be classroom layout, where in a physical school students can be set up individually with desks, around tables or in groups of varying sizes either randomly, based on academics or behavior. In an online video conferencing classroom where spatial distance is hard to replicate, the findings are not going to apply. Online schooling has

seen significant growth with 453,000 students in the US enrolled in virtual schools according to the National Centre for Education Statistics with a forecasted growth to 1,200,000 students by 2024 (William Blair Equity Research, 2019) without considering the effects of COVID-19.

Student academic outcomes are an important factor to be concerned about in education. The need to quantify outcomes, as opposed to qualitative measures of success that are indicative of outcomes, is especially important given the research which challenges the linkage between student outcomes and student satisfaction in various areas of learning (Zurner et al, 2015). It is quite difficult in education to get the support of the necessary stakeholders to deploy large-scale randomized control trials and so there tends to be more cross-sectional analysis and case studies as opposed to randomized control trials in the existing literature. The skeptics of randomized control trials argue that they generally produce no change on student outcomes and in some cases, can produce negative effects but can create significant distraction or resource utilization to implement (Hammersly et al, 2007).

While student outcomes are clearly more important than student satisfaction, student satisfaction trends help us to understand some of the mechanisms that may drive improved student outcomes. On the first order, a student who enjoys learning is more likely to spend time on activities that boost student outcomes. On the second order, a student who is achieving well is likely to feel the euphoria of success and progress which in turn may drive student satisfaction up, which in turn is likely to drive more focused achievement. This reflexive loop between student outcomes and student satisfaction is an interesting relationship to explore. This mechanism is complicated and constantly debated (Zurner at al, 2015). For example, potentially student satisfaction only matters in a binary context in which student outcomes don't improve if student satisfaction is low but student satisfaction being high doesn't necessarily drive student outcomes. Additionally, there may

well be development differences in what drives outcomes for five year old students as compared to eighteen year old students.

2.4 Methodology Used in the Systematic Review

Method for Inclusion/Exclusion

Our inclusion criteria for the drivers of student satisfaction and outcomes in the context of virtual schooling:

1. Availability as a journal article between 1996 and 2018 in English
2. Consider students engaging in K-12 schooling in an online school between ages five to eighteen.
3. In defining student outcomes, focus on academic achievement or retention
4. Only one of student outcomes or student satisfaction needs to be included in order for the research article to be included.

Our exclusion criteria for the drivers of student satisfaction and outcomes in the context of virtual schooling:

1. Availability as a journal article in 1995 or a previous year
2. Availability in a language that isn't English
3. Considers students engaged in university education either at the undergraduate or postgraduate level
4. Considers students engaged in kindergarten or other education prior to commencing primary school

This timeframe was selected because virtual schooling began to emerge as a major trend in the mid 1990s meaning that historical research prior to this time period was limited.

I conducted electronic searches in the following databases:

1. JSTOR
2. ProQuest Education

3. PsychInfo
4. ERIC
5. Ovid
6. EBSCOHost
7. Google Scholar

In total 142 electronic searches were conducted (7 search engines and 2 search criteria) and the information was stored in Covidence, a systematic review management software.

The Cambridge Quality Score Assessment Framework was used to assess the quality of the various articles in order to appropriately weight findings. This framework consists of three criteria: correlate score (measured by assessments of sampling, response rates, sample size, measures of correlate and measures of outcome), risk factor score (based on the type of data collected) and causal risk factor score (based of variation of risk factor and the analysis of change). Murray et al (2009) designed the Cambridge Quality Score Assessment Framework to aid in "identifying high-quality studies of correlates, risk factors and causal risk factors for systematic reviews and meta-analyses". In developing this framework, correlation was defined as variables that have been shown to have an association with one another, risk factors was defined as variables that have a predictive relationship with the outcome because of clear temporal ordering and finally, causal risk factors are defined as risk factors that are variable which cause a shift in the risk for the outcome when they vary (Kraemer at al, 2005).

2.5 The Cambridge Quality Checklists

Correlate score (out of 5)	
Sampling	
1	Total population or random sampling
0	Convenience or case–control sampling
Response rates	
1	Response and retention rates ≥70% and differential attrition ≤10%
0	Response rate <70% or retention rate <70% or differential attrition >10%
Sample size	
1	Sample size ≥400
0	Sample size <400
Measure of correlate	
1	Reliability coefficient ≥.75 and reasonable face validity or criterion or convergent validity coefficient .3 or more than one instrument or information source used to assess correlate
0	None of the above
Measure of outcome	
1	Reliability coefficient ≥.75 and reasonable face validity or criterion or convergent validity coefficient .3 or more than one instrument or information source used to assess correlate
0	None of the above
Risk factor score (out of 3)	
1	Cross-sectional data
2	Retrospective data
3	Prospective data (or study of fixed risk factor)
Causal risk factor score (out of 7)	
1	Study without variation in the risk factor No analysis of change
2	Study with variation in the risk factor but inadequately balanced No analysis of change
3	Study without variation in the risk factor With analysis of change
4	Study with variation in the risk factor but inadequately balanced With analysis of change
5	Study with variation in the risk factor and adequately balanced

	No analysis of change
6	Study with variation in the risk factor and adequately balanced
	With analysis of change
7	Randomised experiment
	Targeting a risk factor

In the first component of the Cambridge Quality Score Assessment Framework, the checklist testing for correlates has five considerations which are binary in nature. This component of the framework is designed to the methodology by which the sample was collected, the response rates and the retention rates of people in the trial, the total sample size achieved and the methodology for assessing how the correlate and outcome were assessed.

In the second component of the Cambridge Quality Score Assessment Framework, the checklist is testing whether a given variable can be defined as a risk factor. In order to assess this, reviewers are asked to assess how the ordering of data in the study occurred. Cross-sectional data, which is the least rigorous, is given a value of 1. Time-ordered retrospective data is given a value of 2. Prospective longitudinal data in which a risk factor is given scores a 3. A key feature of a prospective study is that at the time of enrollment of subjects in the trial, none of the subjects has developed any of the outcomes under consideration.

Of the three components of the Cambridge Quality Score Assessment Framework, the largest contributor to overall score and subsequently the most significant element in the assessment criteria is the causal risk factors. This part of the framework is testing for common issues with causality in non-randomized studies. The first consideration is whether, within-individual changes in a given outcome variable, are also associated with within-individual changes in the specified risk factor. The second consideration is whether the study has controlled for alternative mechanisms to explain the findings. In order to score a full 7/7 in the assessment framework, one must conduct a randomized control trial with a specific targeted risk factor. In

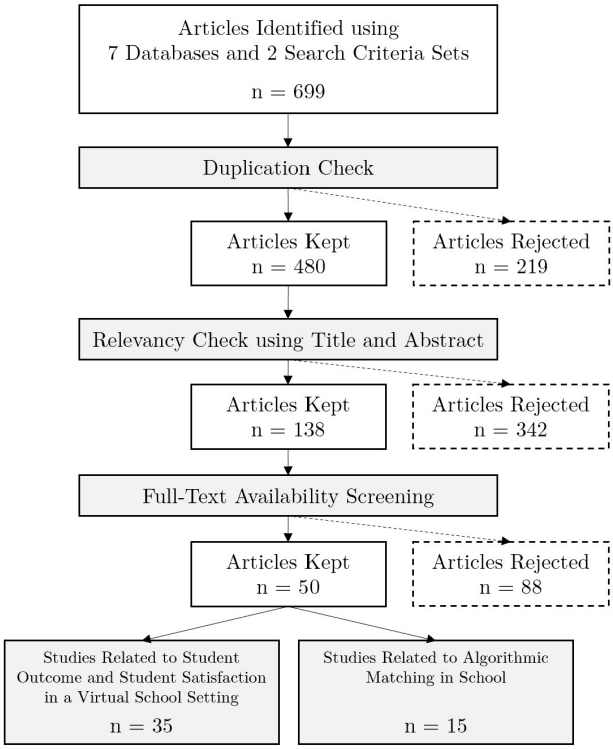


Figure 2.1: Article Search Process

the absence of full randomization, a 6/7 can be scored if variation in the risk-factor is related to within-individual changes in an outcome variable while controlling for all relevant confounding variables.

Alternative assessment frameworks including the Maryland Scientific Methods Scale (Farrington, 2003) was considered, but ultimately in considering the set of research papers, the Cambridge Quality Score Assessment Framework was the most compelling choice. One major benefit of this approach is the appropriate weighting of research methodology for epidemiological studies, with more weight being given to designs from which stronger causal inferences can be made. The Maryland Scientific Methods Scale also lacked a thorough way to deal with temporal risk factors.

The flow chart below outlines the removal process across the two systematic reviews conducted in chapter two and three:

In total, between all search results in all journals across the two systematic

reviews conducted, 699 articles were identified through database searching. 480 articles remained after duplication removal. 342 articles were then removed during title and abstract screening. Of the remaining 138 articles, 8,875 articles were removed during full text screening and 50 articles made it to extraction. Of the 50 articles, 35 research papers related to student outcomes and student satisfaction in virtual schooling.

In total, there were 35 research papers that satisfied the search terms and inclusion/exclusion constraints. Of those, 12 of the research papers focused on factors that impacted student satisfaction in one-to-one online learning. 23 of the research papers focused on factors that impacted student outcomes in 1-1 online learning. 32 of the research papers focused on virtual schooling at the elementary or high school level and three research papers primarily focused on the university level, but with sufficient relevance to the high school segment. It is notable that 33 of the 35 research papers were set in the United States with only two other locations (the Netherlands and Canada) featuring.

The research articles average scores of 6.46 with a standard deviation of 3.14 on the Cambridge Quality Score Assessment Framework.

A priori decision was made to not perform a meta-analysis given the variance in types of articles that were prevalent in the literature, specifically the number of case studies or qualitative research articles which do not lend well to a consolidated quantitative survey.. There was also a wide variety of different outcome measures given the breadth of the search term meaning it was not functionally possible to consolidate the experiments. A systematic review that was going to feature a meta-analysis would require a more specific definition of the "student outcome" or "student satisfaction" measure in question. A proportion of studies would need to have similar measures in order to be able to synthesize in a meta-analysis. A priori decision was made to instead narratively synthesize the data

2.6 Summary Tables of Systematic Review

The table below summarizes the various findings of the systematic review.

Table 1 - Reference: Each research article is matched with an ID to assist with navigation across the tables

	Research Title	Description of Participant Population	Intervention
1	Integrating Data Mining in Program Evaluation of K-12 Online Education	Northwestern US K-12 Students	Program evaluation through analyses of student learning logs, demographic data, course-end surveys in online K-12 program
2	Differences in Student' Satisfaction of the Economics and Personal Finance Virtual High School Course Between Students Attending Economically Disadvantaged and Non-Economically Disadvantaged Schools in Virginia	High School Students Enrolled in an Economics and Personal Finance Virtual/Online High School Course in Virginia	Differences of perceived overall satisfaction of high school students enrolled in a virtual course in different socioeconomic status schools, as measured by the e-Learning Student Satisfaction (ELS) instrument
3	Faculty/Student Mentor Program: Effects on Academic Performance and Retention	University Students (Undergraduates) and faculty	Protégé and Mentor matching based on Gender, Ethnicity, and GPA
4	Distance Education Use in Rural Schools	Postsecondary Students	Survey conducted to determine extent to which distance education is being used by rural schools, technologies used, curriculum areas impacted, perceived needs for distance education, satisfaction with distance education, and barriers to distance education use.
5	Job Satisfaction, Organizational Commitment, and Turnover Intention of Online Teachers in the K-12 Setting	Part-time/Full-time Teachers from 11 K-12 Online Schools in the Southeastern US	Influence factors on K-12 Online teachers' job satisfaction, organizational commitment, turnover intentions
6	Public Online Charter School Students: Choices, Perceptions, and Traits	Students Newly Enrolled in a Statewide Public Online Charter School Program in the United States	Online students' perceptions, preferences, and noncognitive traits in relation to online charter high school
7	Student Perception of Teacher Feedback and the Relationship to Learner Satisfaction in a High School Online Course	Students and Teachers at District-led Online Learning Program in Missouri	Correlation between students' perceptions of the amounts and levels of feedback they received from their instructor and overall course satisfaction

8	Comparison Study: Virtual and Traditional Classrooms on High School Students' Mathematics and English Academic Achievement	Traditional and Virtual Students in High Schools in a Rural Southern School District in Southeastern Region of North Carolina	Impact of virtual classrooms and traditional classrooms on high school students' mathematics and English academic achievement
9	An Investigation into Reported Differences Between Online Foreign Language Instruction and Other Subject Areas in a Virtual School	Secondary Level Students and Teachers	High school students took part in online surveys and responses from foreign language students were compared to students in five other subject areas. Follow-up survey had the purpose of elaborating on responses to first survey.
10	An Analysis of the Virtual Classroom	Syracuse University Students	Impact of enrollment on student evaluations
11	Ethical Behaviours of Student Teachers' Mentors in Forced Same-gender and Cross-gender Matches in a Malawian Initial Primary Teacher Education Programme: Implications for Mentor Selection and Development	Student Teachers in Second Term of Teaching Practice Mentoring Programme	Gender and cross-gender matching
12	The Validation of One Parental Involvement Measurement in Virtual Schooling	Parents of Grades 9-12 Students of Virtual School	Item and factor loading of parental involvement mechanisms
13	Priorities in K-12 Distance Education; A Delphi Study Examining Multiple Perspectives on Policy, Practice, and Research	Practitioners Engaged w/ Operations of Distance Learning, Policy Makers of Distance Education Programs, Researchers of Distance Education	Priorities in a distance education program
14	Students with Special Health Care Needs in K-12 Virtual Schools	Parents of Special Health Care Students Attending Virtual School	Student performance in virtual school and demographics
15	Matching Teacher Feedback and Student Perceptions in a Collaborative Learning Environment	Students and Teachers in Dutch University Prepatory Secondary History Education	Relationship between oral teacher feedback and the students' perceptions of the oral teacher feedback during collaborative learning
16	Investigating Student Satisfaction and Retention in Online High School Courses	Students Enrolled in the CPPS Summer Program in Grades 8-12	Student demographic and academic variables relating to student satisfaction
17	Online Schools and Children with Special Health and Educational Needs: Comparison with Performance in Traditional Schools	Parents of Students in OHS in 3 US States	Basic and applied demographics of US online school users in comparison to student achievement in traditional versus online users + student achievement in traditional versus online schooling environments.
18	Fostering Student Success and Engagement in a K-12 Online School	Parents of Students in OHS	Factors affecting student achievement in K-12 online school

19	Academic Outcomes for North Carolina Virtual Public School Credit Recovery School	Students who Took Credit Recovery Courses Offered by NCVS and All Identifiable Students who did Credit Recovery in Other Methods	Students' academic success in NCVS credit recovery in comparison to students taking credit recovery outside of NCVS (in traditional setting)
20	The Effects of Response to Intervention (RTI) On Student Achievement in a Virtual High School	Students and Staff Members within Virtual High School in Rural County of Tennessee	Effects of RTI on student achievement in VHS
21	Student Performance in Virtual Schooling: Looking Beyond the Numbers	Students in CDLI Program	Student achievement in test scores to determine success in VHS environment at same rate as classroom counterparts
22	Inqiury-based Learning and e-Mentoring via Videoconference: A Study of Mathematics and Science Learning of Canadian Rural Students	Canadian Rural Students	Impact of E-mentoring on student learning
23	Perceived Advantages and Disadvantages of an Online Charter High School	Ohio Parents, Students, and Teachers	Comparison of standardized achievement test scores between students in online, similar, and other schools
24	The Relationship of Learner-Centered Beliefs of North Carolina Virtual Public School (NCVPS) Teachers and Student Achievement on the North Carolina End-of-course Assessments	Teachers from across North Carolina along with the Achievement Levels of their Students	Teachers' beliefs about the learner, learning, and teaching as well as the influence of their beliefs on student achievement in Algebra I, Biology, and English I classes
25	The Nature of an Adolescent Learner Interaction in a Virtual High School Setting	Utah Resident Freshman Students Enrolled in Virtual High School Courses	Effects of interactions in virtual high school (VHS) courses on course outcomes
26	Growth and Performance of Fully Online and Blended K-12 Public Schools	Students Enrolled in Full-time, Public Elementary and Secondary Virtual and Blended Schools in the US	Public and private benefits of online schooling
27	High Enrollment Course Success Factors in Virtual School: Factors Influencing Student Academic Achievement	State-wide Part-time and Full-time Students Enrolled in Virtual Learning Environment	Influences of student level variables on dependent variable: students' final score in online score
28	An Exploratory Case Study of Middle School Student Academic Achievement in a Fully Online Virtual School	Middle School K-12 Students in a Pennsylvania Cyber School	Perceptions of student, teachers, learning coaches and academic achievement
29	Online High School Student Achievement on State-Issued Tests: A Case Study	Georgia Virtual School Admin and Teachers	Characteristics of a successful online school that help determine what may limit student achievement and practices educators can use to increase academic achievement in the virtual classroom
30	Can Virtual School Thrive in the Real World?	Ohio Virtual and Traditional School Students	Virtual students' PI scores in comparison to traditional school counterparts

31	An Online High School "Shepherding" Program: Teacher Roles and Experiences Mentoring Online Students	Teachers in MHA	n/a
32	Creating Virtual Classrooms for Rural and Remote Communities	n/a	n/a
33	The Scaled Arrival of K-12 Education: Emerging Realities and Implications for the Future Education	n/a	n/a
34	Cyber Charter Schools: Evolution, Issues, and Opportunities in Funding and Localized Oversight	US Cyber Charter Schools	n/a
35	Virtual High Schools: Improving Outcomes for Students with Disabilities	American Students with Disabilities	Dropout rates and influences relating to disabilities

Table 2 - A List of Varying Studies, Outcomes, Settings, Participants, Control Groups, and Findings

	Comparator	Outcomes	Setting	Participants	Assumed Control Group Risk and Corresponding Intervention Group	Mean Difference or Standard Mean & Confidence Interval
1	Clustering Analysis and Decision Tree Analysis	Demographic, engagement, satisfaction and performance	Students in Northwestern State US K-12 Online Institution	7539 students	n/a	Electives Subjects pass rates for female vs male: $p<0.05$ (Fall 2009, Spring 2010 semester)
2	MANOVA, ELS Instrument, Kolmogorov-Smirnov Test	Did not find a statistically significant difference in perceived overall satisfaction, content, personalization, and learning community, but found a significant difference regarding learner interface	Urban/Suburban Virginia High Schools	249 students	Overall perceived satisfaction: mean: 3.94 sd: .793 Content: mean: 3.96 sd: 1.15 learner interface: mean: 4.10 sd: 1.06 personalization: mean: 3.59 sd: .694 Learning community: mean: 3.7 sd: .774	Assumption for normality was not found tenable at the .05 alpha level for each dependent variable in group 1 (economically disadvantaged school)
3	ANOVA, Chi-square	Higher GPA for mentored students (2.45 vs 2.29), more units completed per sem. (9.33 vs 8.49), & lower dropout rate (14.5% vs 26.3%)	Metropolitan University on US West Coast	678 students and faculty	GPA Variable 9 (1st sem.): protégé: 2.50 SD:0.93, control: 2.20 SD: 1.11, Gender Variable: female mentor contact: 7.95, male mentor contact: 6.50, Ethnicity: Latino: 2.51 SD:0.82, African American: 2.23 SD: 0.80, Nat. American: 2.66 SD: 0.66, Other: 2.56 SD: 0.78	GPA: 0.3, $p<0.001$, Gender: 1.45, $p<0.01$, Ethnicity: none
4	SPSS Analysis	Positive support for distance education, majority of districts felt a need for distance education	All districts in the US that are REAP Qualified (Rural Education Achievement Program)	417 school districts	Taken D.E course: 12.15 SD: 16.18, Prev. used D.E: 5.41 SD:0.71, Currently Using: 13.99 SD: 17.11, Completion of prev. used: 87.67 SD:26.71, Completion of currently using: 89.39 SD: 24.15	No significant difference $t(320)=-0.47$, $p>.05$
5	Quantitative and Qualitative Analysis (Cross-sectional Survey Design, Likert Scale)	Positive student-teacher relationship, input in planning of curriculum, meeting students' needs were main factors attributing to teaching satisfaction	Southeastern K-12 Online schools	108 online teachers	Student Interactions: mean: 3.14 sd: 4.04, Affordance: mean: 4.4 sd: 2.69, Inst. Support: mean: 4.85 sd: 3.59, Courseware and Instr: mean: 3.26 sd: 3.81, Overall Satisfaction: mean: 3.83 sd: 2.38	Highly satisfied: Affordance: $p<0.001$, Inst. Support: $p<0.01$, Overall satisfaction: $p<0.001$

6	SEM (Structural Equation Modeling), ANOVA	Online charter school will become important in education system because : It may serve the student who could not or would not enroll in other options, it may cause mediocre or bad schools to improve their education service quality in order to stay competitive in the ecosystem.	US Online Charter Schools	524 respondents	(Group A - Online Discussion is helpful) Prior online learning experience: mean: 1.60 sd: .907, Working in groups: mean: 3.02 sd: .818, Difficult with traditional materials: mean: 3.6 sd: .883, Difficulty with traditional schooling: mean: 3.51 sd: .653 ... (Group B - Online Discussion is not helpful) Prior online learning experience: mean:1.45 sd: .845, Working in groups: mean: 2.65 sd: .946, Difficult with traditional materials: mean: 2.04 sd: .994, Difficulty with traditional learning: mean: 3.25 sd: .796	Prior learning experience: mean difference: .151, p=0.071, Working in groups: mean difference: .376, p<0.001, Difficult with traditional materials: mean difference: .317, p<0.001, Difficulty with traditional schooling: mean difference: .264, p<0.001 ...
7	(Inferential Stats Techniques) Pearson Correlation Coefficient, Likert-scale, T-tests for r	stronger correlation between students' perceptions of the amount of feedback they received and overall course satisfaction than the level of feedback they received.	District-led online learning program in one large accredited district in Missouri	83 students and 6 teachers	(Course A) t-test for r = 0.557>0.05, (Course B) t-test for r = .021<0.05, (Course C) t-test for r = 0.310>0.05, (Course D) t-test for r = 0.400>0.05	A: not significant of a relationship between 2 variables, B: null reject and alt. hypothesis considered as relationship between 2 variables exist at significant level, C: no significant relationship between 2 variables, D: no significant relationship between 2 variables
8	ANOVA, T-tests of Independent Means	Virtual classroom instruction and the experience in high school is beneficial to students and is as effective as traditional classroom instruction	High Schools in a Rural Southern School District in Southeastern Region of North Carolina	70 virtual school students and 70 traditional school sample	(Virtual) mean score: 244 std:7.6 (traditional) mean score: 242 std: .4	Diff: 2.0, t-value: 1.53, prob level: 0.22
9	One-way ANOVA, SPSS Software, REGWQ Test	Foreign language students had significantly lower perceptions of online courses	North Carolina Virtual Public School	559 students and 32 teachers	Q2 %agree/strongly agree (5-point Likert scale) mean - English:3.42, Career: 3.19, Soc. Studies: 2.80, Science: 2.92, For. Language: 2.45, Math: 2.61	Significant at the 0.05 level
10	Regression Analysis, Scatter Charts, Fixed-effects regression analysis	Higher enrollments resulted in lower teaching evaluations	Syracuse University School of Information Studies	n/a	Regression coefficients for Constant 4.4041, Enrollment -0.0190	Significant at the 0.05 level

11	Cross-sectional Study	Ethical and unethical behaviors were prevalent	Teacher Training Colleges in Malawi	616 student and teachers	Control: mean:2.6 sd:0.59 vs Treatment: mean:2.7 sd:0.57	Not significant at 5% significance level
12	Confirmatory Factor Analysis	Validation of parental involvement instrument	State-level Institution in Southeastern US	938 parents	Reliability coefficient of 4 scales (parent report of encouragement, modeling, reinforcement and of instruction): (0.91, 0.88, 0.90, 0.93); GOF indices CFI=.973, TLI=.972	x^2 (11064.904) significant (p<0.001)
13	Delphi Method, Kruskal-Wallis H Test	Most important priorities were evaluation of course design and delivery, best practice, & accountability	12 States in the US	29 respondents	Evaluation of course design and delivery: (R3) mean: 2.62 sd:1.75 (R2) mean: 4.07 sd: 2.06, best practice: (R3) mean: 3.65 sd: 2.43 (R2) mean: 4.11 sd: 2.68, accountability: (R3) mean: 3.81 sd: 1.92 (R2) mean: 4.33 sd: 2.34	No significant differences between subgroups for any of the individual statement importance ratings (p>0.05)
14	CSHN	Study #1: male African American students with SHCN has significantly lower grades in comparison to female population of other ethnicities and healthy children - more likely to have lower grades online than in traditional. Study#2: no such differences.		290 parents	students w. SHCN - 'usual' (traditional schooling) grade: x bar (95% CI) = 2.92, virtual grade: x bar=3.38. students w/ no SCHN: - 'usual' grade: x bar=3.23, virtual grade: x bar=3.50	p<.01 for all
15	MSLQ (Motivated Strategies for Learning Questionnaire), SAFL-Q (Student Assessment for Learning Questionnaire), KMO (Kaiser Meyer-Olkin measure), MANOVA, Fisher's Exact Test	Relationship between the feedback quality and the students' perceptions of the feedback quality was found to be very weak	Dutch University Prepatory Secondary History Education	77 students, 2 teachers	first model: 5 predictor variables together explain significant part of students' perception of elaborative teacher feedback R^2=.16, F(5,65)=2.44, second model: explains a significant part of students' perception of the dependent variable R^2=.12, F(3,67)=3.14	First model: p=.044, second model: p=.031
16	Independent T-tests, Hedges' g for Calculation of Effect Sizes	Students repeating courses, students in electives, and females reported higher rates of satisfaction than reported by their peers	CPPS Summer School 2013 Program	71 participants	(Earned course credit) mean diff: 0.003, std. error diff: 0.072, (grade earned) mean diff: 4.118, std. error: 5.889, ...	No significant differences w/ conditions: t(301)=-0.044, p=.965, t(301)=.699, p=0.485

17	Multivariate Regression Techniques	Prevalence of CSHCN was high in online schooling, Children who were male, Black, or had special health care needs reported significantly lower grades in both traditional and online schools	State-led Online Schools in Southeastern Region of US	1971 parents	adjusted odds ratio [aOR] 1.45, 95% confidence interval [CI] 1.29–1.62 for CSHCN, P < .001? aOR 2.73, 95% CI 2.11–3.53 for black children, P < .001	Significance at p<0.001
18	Theory Generated Coding, Semi-structured Interviews	Student success and parental guidance	Online High School in Western United States	4 sets of parents	Unsuccessful: frequency code 'parent monitoring' (9%) Successful: Frequency code 'parent monitoring' (14%)	n/a
19	Multivariate Least Squares Regression Model, Multivariate Logistic Regression Models.	NVCS students were less likely than other credit recovery students to be economically disadvantaged, greater proportion entered HS proficient in math + reading. Small difference in short term success rates between NCVS and other CR students.	North Carolina Virtual Public School	n/a	Demonstration of proficiency on North Carolina end-of-course exam retest, all courses combined: (NCVPS CR student) odds ratio: 0.6119 standard error: 0.0273	Significant at p<0.001
20	School's Curriculum and Intervention Platform; Edmentum-PLATO, and the District's Benchmark Assessment	Treatment group's posttest scores significantly increased in comparison when compared to group w/ both intervention assessment methods	Virtual High School within Rural County of Tennessee	112 students and 7 staff members	Pair 1 (EDMATHRPE) mean: 50.82 sd: 17.8 (EDMATHPOST) mean: 55.26 sd: 16.183, Pair 2 (EDELARPE) mean: 57.57 sd: 13.416 (EDELAPOST) mean: 62.12 sd: 12.412	interaction between 2 variables was significant: F(1,110)=37.99, p<0.001
21	Descriptive Statistics	When distance education is accessible to student, they score lower than their classroom counterparts	CDLI Program in Canada	n/a	Average student performance: (rural) public exam:61.93 final course: 68.52, (Urban) public exam: 62.76 final course: 67.85, (web-based) public exam: 62.22 final course: 68.86, (classroom) public exam: 2.41 final course: 68.14	No performance differences between web-based or virtual school students and classroom-based or brick-and-mortar students in both the public examination and final course scores in the first four years of CDLI data
22	Descriptive Analysis	IBLE model enhanced students' learning	Rural School in Canada	67 students	Control: mean: 85.4 sd: 24.54 vs Treatment: mean:97.35 sd: 28.55	No significant difference at p=0.056, t(65)=59.03
23	Focus Groups	Online charter school studied provided flexibility and individualization for student instruction	Ohio	44 students, parents, and teachers	Online Charter School averaged: 60%, Similar school comparison averaged: 49.6%	

24	ANOVA, Independent Sample T-tests, Cronback's Alpha, Pearson's Product -Moment Correlation Coefficients	No statistically significant relationship found between non-learner-centered beliefs and student achievement + learner-centered beliefs and student achievement	North Carolina Virtual Public School	31 teachers	LCB (Level I/II): mean: 3.1071 sd: 0.35191, (LCB Level III/IV): mean: 3.1238 sd: 0.46641	Not significant at the $p<0.05$ level of significance
25	ANOVA, Spearman Rho Correlation Analysis	Higher correlations with course outcomes for learner-learner interactions	VHSU (Virtual High School of Utah)	250 students	Significant interaction between students' grade and time spent on learner-learner interaction: ($r=0.257$, $p=0.020$)	n/a
26	Data Visualization & Exploratory Data Analysis / Calculation of Mean Scale Scores and Achievement Levels based on Subject Area Results	More than 1/2 of student populations that attend virtual and blended schools are in schools operated by private for-profit EMOs (education management organizations) / Enrollments in charter virtual and blended schools are substantially larger than district schools.	Virtual and Blended Schools in US	n/a	Teacher student ratios - all virtual schools: mean: 35.03 sd: 36.43	n/a
27	ANOVA (RA model), HLM Technique	Time spent in the LMS was most significant factor in student's achievement	State-Level Virtual High School in Midwestern US	1794 students	ICC of Algebra 1 (1): 0.32, Algebra 2 (1): 0.46, Geometry 1: 0.32, Biology 1: 0.08, English 1 (1): 0.33, English 2 (1): 0.60, American History 1: 0.30, American Gov.: 0.24	n/a
28	Archived Record Review, Directed Observation, Interviews, Document and Artifact Analysis, Focus Group Interviews	Student participants share learner characteristics in common, teachers play an important role in the virtual school model, and specific aspects of parental involvement were revealed that indicate a learning coach's approach may promote academic achievement	K-12 Cyber School in Pennsylvania	8 students and 8 learning coaches, and between 9 teachers for online focus group	n/a	n/a
29	Interviews of OHS Teachers, In-depth Semi-structured Interviews of OHS Admin, Focus Groups of OHS Teachers	The importance of a quality online curriculum, highly skilled and motivated educators, individualized instruction and feedback, and timely and consistent communication between the teacher and all parties invested in the learning process positively impact online student learning	Georgia Virtual School	13 educators from Georgia Virtual School	n/a	n/a

30	PI Scores	Ohio's virtual schools have grown rapidly, but have also experienced much lower levels of school performance than traditional schools	Ohio Virtual School	n/a	n/a	n/a
31	Case Study and Focus group	Increase of job satisfaction for teachers	Mountain Heights Academy	5 OHS instructors, 56 on-site facilitators	n/a	n/a
32	n/a	n/a	n/a	n/a	n/a	n/a
33	n/a	n/a	n/a	n/a	n/a	n/a
34	n/a	n/a	n/a	n/a	n/a	n/a
35	n/a	n/a	n/a	n/a	n/a	n/a

Table 3 - A list of Quality Scores and Quality Score Totals.

	Cambridge Quality Checklist Quality Score			
	Correlate Score	Risk Factor Score	Causal Risk Factor Score	Total
1	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	11
2	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	6: Study with variation in the risk factor and adequately balanced, with analysis of change	10
3	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 1	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	9
4	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 0	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced, with no analysis of change	9
5	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	9
6	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 1	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	9
7	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	1: Cross-Sectional Data	4: Study with variation in the risk factor but inadequately balanced with analysis of change	9
8	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	9
9	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective	5: Study with variation in the risk factor and adequately balanced with no analysis of change	9
10	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	9
11	Sampling: 0, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	1: Cross-Sectional Data	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	8
12	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 0	2: Retrospective	2: Study with variation in the risk factor but inadequately balanced with no analysis of change	8

13	Sampling: 1, Response Rate: 0, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 1	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	8
14	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	8
15	Sampling: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	4: Study with variation in the risk factor but inadequately balanced with analysis of change	8
16	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	1: Cross-Sectional Data	5: Study with variation in the risk factor and adequately balanced with no analysis of change	8
17	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	4: Study with variation in the risk factor but inadequately balanced with analysis of change	8
18	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 1	2: Retrospective	4: Study with variation in the risk factor but inadequately balanced with analysis of change	8
19	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 1	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	8
20	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 1	1: Cross-Sectional Data	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	7
21	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	7
22	Sampling: 0, Response Rate:1, Sample Size: 0, Measure of Correlate: 0	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	6
23	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 1	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	6
24	Sampling: 1, Response Rate: 0, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	6
25	Sampling: 1, Response Rate: 0, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective	3: Study without variation in the risk factor but inadequately balanced, with analysis of change	6
26	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	2: Retrospective	1: Study without variation in the risk factor and no analysis of change	6

27	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective	2: Study with variation in the risk factor but inadequately balanced with no analysis of change	5
28	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective	1: Study without variation in the risk factor and no analysis of change	5
29	Sample: 1, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 0	1: Cross-Sectional Data	1: Study without variation in the risk factor and no analysis of change	5
30	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 0	2: Retrospective	1: Study without variation in the risk factor and no analysis of change	3
31	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 1	0: No Data	0: Theoretical Study	0
32	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 0	0: No Data	0: Theoretical Study	0
33	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 0	2: Retrospective	0: Theoretical Study	0
34	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 0	0: No Data	0: Theoretical Study	0
35	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate 0, Measure of Outcome: 0	0: No Data	0: Theoretical Study	0

2.7 Key Findings of Research Papers In the Systematic Review

In this section, I summarize the key insights from each individual article. This is designed as an exhaustive list of all papers in the study to provide readers with a consumable executive summary of the literature. In "Discussion of Key Themes Emerging from Systematic Review," I summarize the consistent themes across all the papers and offer a deeper discussion of findings which is more useful for most audiences.

The Nature of an Adolescent Learner in a Virtual High School Setting (Borup et al, 2013) found that time spent on peer-to-peer interactions between students in the same class had a statistically significant impact on students' grades. The more time spent on peer-to-peer interactions, the higher the students' grades achievement was in the sample.

Integrating Data-Mining in Program Evaluation of K-12 Online Education (Hung et al, 2012) used an analysis of student learning logs, demographic data and end-of-class surveys to generate deeper understanding of students' shared characteristics. Decision-tree analysis was used to create a predictive model for student outcomes and student satisfaction if other individuals took a given class. The paper also argued that there is a lack of confirmed relationship between student performance and satisfaction with leading papers showing a combination of positive and neutral results.

Inquiry-Based Learning and E-Mentoring via Videoconference: A Study of Mathematics and Science Learning of Canadian Rural students (Li et al, 2010) attempted to use an inquiry-based learning model to compare student learning outcomes in a rural context. Some relationships with improving student motivation, understanding and career awareness were found but the paper was narrowly

statistically insignificant overall at the 5% significance level.

An Analysis of the Virtual Classroom (Kigma et al, 2006) found a statistically significant positive relationship between higher enrollment rates and lower student evaluations of the teacher at the university level. It is important to note that this could be caused by psychological bias in perception by the student even if there is no impact on student outcomes but is significant regardless. This may relate to a "perceived value" effect where students put a premium on smaller class sizes because of the perception of a more personalized experience, even if it doesn't create any difference in the learning outcome.

Fostering Student Success and Engagement in a K-12 Online School (Curtis et al, 2015) used semi-structured interviews with four sets of parents to draw a relationship between parent monitoring and student success. A higher proportion of references to parent monitoring were found in the student cohorts regarded as "successful" but the sample size was too small for any statistically meaningful conclusions to be reached.

The Validation of One Parental Involvement Measurement in Virtual Schooling (Liu et al, 2010) was designed to validate the applicability of a specific scale that could be used as a proxy for parent involvement. The paper states the clear empirical relationship between parental involvement and student achievement but notes the wide variety of measurements used to capture parent engagement. As such, the scale is proposed and is found to be statistically significant in a sample of 938 parents.

High Enrollment Course Success Factors in Virtual Schools: Factors Influencing Student Academic Achievement (Liu et al, 2011) found that a student's time spent in the online learning management system was the most significant predictor on an improvement in the student's online grades. It is unclear if the online learning management system is the driver of improved outcomes directly or rather heavy engagement is a strong proxy for a talent or engaged student that is also spending substantial time studying the course material.

Creating Virtual Communities for Rural and Remote Communities (Rao et al, 2011) does not perform any statistical analysis and was given a Cambridge Quality Score Assessment of 0. It does focus on advancing the argument that it is important to leverage video-conferencing solutions and technology features that enable students to collaborate with others, especially for those learners from indigenous communities to replicate their traditional learning experiences with peers and mentors. While the paper makes interesting points, the evidence is noticeably sparse.

An Investigation into Reported Differences Between Online Foreign Language Instruction and Other Subject Areas in a Virtual School (Oliver at al, 2012) considered a sample of 559 students and 32 teachers and found that language classes are perceived to be worse online than other subjects which was significant at the 5% level. The primary driver was the lack of being in an environment with physical classmates for the purposes of discussion which isn't as significant in other subjects such as mathematics. A follow-up survey with open-ended questions was provided and responses from 119 students and 19 teachers found that modifying specific aspects of teaching, fostering increased collaboration and providing adequate support for student learning needs could help to offset the drawbacks of online learning.

Perceived Advantages and Disadvantages of an Online Charter School (Shoaf et al, 2007) performed a series of focus groups with 44 students, parents and teachers to try and elicit insights into community perceptions of the strengths and weaknesses of virtual schooling. The most consistent themes were virtual schooling provided a more appropriate pace of instruction, provided greater access to special subjects that were difficult to find in traditional brick-and-mortar environments, offered greater access to individualized instruction, facilitated easier modification of lessons, enhanced flexibility in structuring the school day but resulted in limited social engagement.

An Online High School "Shepherding" Program: Teacher Roles and Experiences Mentoring Online Students (Drysdale et al, 2014) didn't perform any meaningful

statistical analysis but conducted case studies and focus groups with five online high school facilitators and 56 on-site facilitators. The primary themes found in the case study was that the online environment allowed teachers to build relationships with students through non-traditional communication channels that are not used in brick and mortar schooling. The primary examples of this would be Google-chat messages, text messages and voice recordings. The recorded nature of this communication and physical barrier between the student and the teacher because they aren't co-located helps to reduce some types of child safety risks and subsequently facilitates more of this informal communication with more relative safety than the offline world. Online environments still present risk of cyber-bullying or other types of harassment which require their own child safety interventions.

Faculty/Student Mentor Program: Effects on Academic Performance and Retention (Campbell et al, 1997) found a statistically significant higher GPA, more units completed per semester and lower drop-out rates for those students with an assigned mentor at a university on the US West Coast. Given online learning has a higher drop-out rate than traditional brick-and-mortar schooling, these effects are likely to be especially important to consider in optimizing the environment.

Distance Education Use in Rural Schools (Hannum et al, 2009) interviewed 419 school districts to better understand their attitude to online learning. Of the 328 who responded, 302 described their attitude to online distance education as either very satisfied or somewhat satisfied. 258 of the 302 continued to use distance education. Of those who stopped, the primary reasons cited were lack of interest or participation by student, time or scheduling issues, lack of support personnel, money, equipment or infrastructure, no longer needed or having hired a qualified in-person teacher.

The Scaled Arrival of K-12 Education: Emerging Realities and Implications for the Future of Education (Bashem et al, 2013) calls for a reduced focus on deploying technology solutions proactively into virtual schools but rather focusing

on instructional goals and design principles. The article offers a theoretical argument about a number of risk factors associated with technology deployment that isn't carefully calibrated for students. This paper received a Cambridge Quality Score assessment of 0 given the lack of empirical analysis but provides interesting context.

Job Satisfaction, Organizational Commitment and Turnover Intention of Online Teachers in the K-12 setting (Larkin et al, 2015) analyzes a sample of 108 online teachers from some Southeastern K-12 online schools and finds the driving factors behind teacher retention are affordance, institutional support and overall satisfaction at the 5% significance level. The paper finds that positive student-teacher interactions, input in planning of curriculum and meeting students needs were main factors attributed to teaching satisfaction. This paper was relatively rigorous for a cross-sectional analysis with a Cambridge Quality Score assessment of nine.

Cyber Charter Schools: Evolutions, Issues and Opportunities in Funding and Localized Oversight (Ellis et al, 2008) is a theoretical paper that makes an argument that because cyber-charter schools are not revenue neutral but rather are a source of funding allocation in education budgets and are growing rapidly, more rigorous standards need to be put in place to calculate funding formulas. The paper argues for more intense student attendance requirements which help to reduce the risk of fraud by charter schools but also impose more rigidity on students in virtual schooling which may reduce their flexibility and happiness which is one of the primary attractions of this schooling option.

Priorities in K-12 Distance Education: A Delphi Study Examining Multiple Perspectives on Policy, Practice and Research (Rice et al, 2006) attempted to quantify some of the most important priorities that drive student outcomes and suggested evaluation of course design and delivery, best practice and accountability, however, with a sample size of only 29 respondents none of the findings were statistically significant at the 5% significance level.

Students with Special Health Care Needs in K-12 Virtual Schools (Fernandez et al, 2016) ran two studies with a total sample size of 290 parents. In the first study, male African American students with special health care needs had lower grades at the 5% confidence level in comparison to the female population of other ethnicities and healthy children and are likely to perform worse in online learning environments than brick-and-mortar environments. In the second study, the effect was not statistically significant. This paper is significant because high-risk students, with or without disabilities or choosing virtual schooling at present and so if these students systematically underperform it would be an important finding.

Academic Outcomes for North Carolina Virtual Public School (NCVPS) Credit Recovery (Stallings et al, 2016) found that NCVPS students were less likely to be economically disadvantaged, and that a greater proportion entered high school proficient in mathematics and English. It also found no statistically significant difference between short-term success rates of NCVPS students and other credit recovery programs. Students were less likely to graduate but those who did were more likely to stay on track for graduation in four years than other students.

Growth and Performance of Fully Online and Blended K-12 Public Schools (Gulosino et al, 2017) performs a broad analysis of virtual schooling in 35 US states. It uses data visualization and exploratory data analysis (Cambridge Quality Score Assessment of 6) to reach a conclusion that virtual and blended schools are generally poorly performing but are seeing increasing enrollments primarily from schools run by for-profit education management organizations. While the methodology isn't particularly robust, the paper purports to offer conceptual and empirical implications of the growing adoption trend on virtual schooling balancing public and private considerations.

Public Online Charter School Students: Choices, Perceptions and Traits (Kim et al, 2012) considers a sample of 524 learners and attempts to understand the

drivers behind strong growth in enrollment in the US using structural equation modelling and ANOVA. The paper finds some of the key drivers underpinning the enrollment growth is prior learning experiences with brick-and-mortar schooling and difficulty with traditional schooling. The paper also argues that online schooling drives competitive pressure higher and forces underperforming brick-and-mortar schools to improve their standards which is an interesting perspective shared by many in the school-voucher community in the United States.

Student Performance in Virtual Schooling: Looking Beyond the Numbers (Barbour et al, 2009) finds no performance difference between web-based, virtual-school students, classroom-based or brick-and-mortar students at the 5% significance level in both the public examination and final course scores in four years of Centre for Distance Learning and Innovation Data.

Virtual High Schools: Improving Outcomes for Students with Disabilities (Liu et al, 2010) theoretically explores the potential for virtual schools to enhance student achievement and completion rates for those with disabilities by optimizing around a framework described as the Five Cs including connect, climate, control, curriculum and caring community. The paper offers a thought-provoking framework but the lack of data resulted in a 0 in the Cambridge Quality Score Assessment framework.

Investigating Student Satisfaction and Retention in Online High School Courses (Rogers et al, 2014). Students were split into four groups including students who were sitting the course for the first time and earned credit, students who were sitting the course for the first time who did not earn credit, students who repeated the course who then earned credit and students who repeated the course who didn't earn credit. No statistically significant demographic differences were found between the students in each cohort. Student survey and interview responses showed higher student satisfaction for enrollment in electives, students who earned credit, students who repeated courses, and female students. Students who earned credit

were described as self-motivated, investing considerable time into their courses. Conversely, students who did not earn credit did not accept personal responsibility for their learning and had difficulty with course pacing. The paper suggested that helping students develop effective learning strategies raised their completion rates for online courses as well as their satisfaction.

An Exploratory Case Study of Middle School Student Academic Achievement in a Fully Online Virtual School (Wolfinger et al, 2016) analyzed eight students, eight learning coaches and nine teachers from an online focus group. The study suggested similarities in learning characteristics between the students, teachers continue to play an important role in the virtual schooling model, the learning management system was perceived as being important to the students and that connection with peers through extra-curricular activities impacted student achievement and motivation in virtual schooling.

The Relationship of Learner-Centered Beliefs of North Carolina Virtual Public School (NCVPS) Teachers and Student Achievement on the North Carolina End-of-course Assessments (Malave et al, 2012) analyzed 31 teachers and the achievement data of their students and found no statistically significant relationship between non-learner-centered beliefs and student achievement or learner-centered beliefs and student achievement using Moment Correlation Coefficients which were not significant at the 5% significance level.

Online High School Student Achievement on State-Issued Tests: A Case Study (Gifford et al, 2017) interviewed 13 educators from Georgia Virtual Schools using in-depth semi-structured interviews and found that a quality online curriculum, highly skilled and motivated educators, individualized instruction and feedback as well as timely and consistent communication were important for success in online schooling. While this paper was qualitative, the methodology was fairly rigorous and well-specified and while the Cambridge Quality Score framework penalizes it

quite heavily, it was quite useful for a case-study paper.

Student Perception of Teacher Feedback and the Relationship to Learner Satisfaction in a High School Online Course (Lemmon et al, 2014) analyzed a district-led online learning program in Missouri with 83 students and six teachers. Across four courses, three of them had no statistically significant interaction at the 5% significance level but one did, suggesting that there is a stronger correlation between students' perception of the amount of feedback they receive and overall course satisfaction than the level of feedback they received.

Differences in Student' Satisfaction of the Economics and Personal Finance Virtual High School Course Between Students Attending Economically Disadvantaged and Non-Economically Disadvantaged Schools in Virginia (Smith et al, 2016) analyzed 249 students and found no statistically significant difference in perceived overall satisfaction, content, personalization or learning community but found a statistically significant difference regarding learner interface.

Online Schools and Children With Special Health and Educational Needs: Comparison With Performance in Traditional Schools (Thompson et al, 2012) tried to establish which format of schooling was more effective for what demographics of children. The paper considered a sample of 1,971 parents and found the prevalence of children with special health and educational needs was higher in virtual schooling and that achievement of these children and black children was lower at the 5% significance level than brick-and-mortar schooling. Additionally, the paper found that parents with a bachelor's degree or higher reported significantly higher online school grades than traditional schooling suggesting the impact of the parent is relatively more significant in virtual schooling.

Can Virtual School Thrive in the Real World (Wang et al, 2014) analyzes the PI scores of students in virtual schools in comparison to brick-and-mortar schools and argues that there is a significantly lower level of school performance. The paper isn't

robust in addressing causation, correlation concerns in controlling for the quality of students enrolling in virtual schools (which are disproportionately credit recovery students) and scored a 3 on the Cambridge Quality Score Assessment Framework.

Comparison Study: Virtual and Traditional Classrooms on High School Students' Mathematics and English Academic Achievement (Tyndall et al, 2014) compares the achievements of 70 virtual school students and 70 traditional school students using ANOVA. The research found no statistically significant differences between student academic achievement in mathematics and English in the two groups. It also found no statistically significant effect of gender or ethnicity on student achievement in either group. The notable exception was female students on final exams, who performed better at the 5% significance level in the traditional classroom. This study suffered from a fairly small sample size.

2.8 Discussion of Key Themes Emerging from The Systematic Review

This systematic review highlights a number of important themes that are relevant to our research. These include (1) the impact of parental involvement, (2) the importance of self-motivation, (3) the impact of peer effects, (4) the impact on general student achievement and (5) the importance of mentorship in online learning environments.

The Impact of Parental Engagement on Student Achievement

On the impact of student achievement, Fostering Student Success and Engagement in a K-12 Online School (Curtis et al, 2015) and The Validation of One Parental

Involvement Measurement in Virtual Schooling (Liu et al, 2010) emphasized the need for a highly motivated parent to participate in the schooling experience as a guardian to maximize the chance of success for a child. This could relate to a number of the findings that disadvantaged students tend to underperform in virtual schooling (Thompson et al, 2012) if highly motivated, engaged parents in education are indeed correlated with more highly advantaged ethnic groups and higher income families. A variety of explanations could explain this relationship that does not link parental engagement to parent income. Children who are excluded from schools with lower socioeconomic backgrounds may opt into virtual schooling meaning the individuals sorting into this delivery mechanism are less likely to succeed from inception. Children with high levels of internalizing symptoms may drop out from school and use virtual schooling. This assumes that children with internalizing symptoms perform worse than average students which is supported in the literature (Hass et al, 2005), but is still somewhat contested. Another simple explanation would be that parents from lower income families cannot afford more costly virtual schools, which may be weaker institutions overall. More highly motivated parents are also likely to have stronger beliefs as to the importance of education, higher aspirations in their own lives which may improve role-modelling, more time available to supervise kids over leisure activities and genetic predispositions towards academic achievement which requires further discussion.

Students from schools such as Florida Virtual may participate dramatically fewer hours per week than a typical student in an actual group-class environment and spend proportionately more time consuming digital content. This enhanced flexibility offered to students is potentially dangerous in that the minimum possible consumption of education delivery is substantially lower in the virtual world. A highly engaged parent acts as a constraint on this consumption falling too low and establishes norms in the home environment around learning that may drive a student to pay more attention and be more accountable for their education. Additionally,

the engaged and more educated parent can act as an impromptu tutor who can remove some of the frictions to engaging with teachers virtually for a student who is confused by a problem as well as providing emotional and other types of support. This is notable because in a traditional school, the friction to asking a question in a classroom is fairly low as a student can simply raise their hand or ask a friend. In a virtual school, depending on the structure, the student may have to log in at a pre-specified time to an "office hours" to ask their question. One could argue that the virtual world reduces friction to asking for help by providing anonymous features and minimizing the negative social stigma of asking for help in front of peers but these effects do not appear to outweigh the previously discussed factors.

The Impact of Self-Motivation on Academic Outcomes

The systematic review highlighted that student self-motivation is correlated with greater success in virtual learning. Inquiry-Based Learning and E-Mentoring via Videoconference: A Study of Mathematics and Science Learning of Canadian Rural students (Li et al, 2010) notes the importance of students taking ownership over their educational journey. Liu et al (2011) found that stronger academic grades are correlated with time spent on the online learning management system the virtual school is run through. This is a proxy for self-motivation because all students need to complete a baseline level of activity on the learning management system (often involving logging into a certain number of classes or submitting a certain number of assignments) but marginal time spent above this threshold is at the student's discretion. As such, spending more voluntary time learning is likely to be positively correlated with student motivation as well as potentially committed parents (who may push the child to spend more time on the learning platform). Additionally, self-motivation is likely to be powerful in capitalizing on some of the strengths of virtual learning. Perceived Advantages and Disadvantages of an Online Charter School

(Shoaf et al, 2007) argues that some of the benefits of virtual schooling is access to more subjects, more personalized instruction and the ability to accelerate more easily. An unmotivated student, seeking to minimize time spent on education, will tend to take fewer niche subjects, will not be so invested in thinking about how to re-calibrate their schooling experience to be more personalized to them and is more likely to stick to whatever has been suggested as default. Additionally, they will not leverage the ability to accelerate which is a strong source of personal fulfilment and motivation (Rogers et al, 2014). In contrast, a self-motivated student will be able to take classes that a brick-and-mortar school may be unable to facilitate such as Latin or computer science given teacher resourcing constraints and accelerate their learning faster than schools can typically cater to because of class-size and scheduling constraints. The student is also likely to be more proactive in optimizing their learning environments, leveraging office hours and taking part in extracurricular activities which are correlated with higher achievement (Rogers et al, 2014).

Peer-Effects Play a Significant Role in Academic Outcomes in Schooling

Peer effects play a significant role in achievement in virtual schooling (Borup et al, 2013). The quality of peers is an input factor in the quality of the overall education system. More intelligent, more engaged peers will increase the quality of discourse in class, the average engagement of students, the pressure on the teacher to prepare prior to lessons and provide a valuable resource who can answer questions in lower-pressure formats than asking teachers. Additionally, peers can relate to other peers and competitive, social dynamics to fit in and be accepted will often drive convergence in academic behavior within groups as similar study schedules, norms around academic achievement, sleep patterns and class selection are pursued. This is explored qualitatively in *Creating Virtual Communities for*

Rural and Remote Communities (Rao et al, 2011). The significance of peers is a very important consideration in designing virtual schools because the role they play varies substantially at present in the existing models. Schools such as Stanford Online High School teach primarily through group instruction and a high proportion of learning hours in a week are with peers and a teacher. Schools that use this type of model tend to be higher cost as they have to pay for increased teacher hours, often with lower class sizes than traditional brick-and-mortar schools and offer substantial flexibility in class selection. As a result, if peer effects are very important, these types of schools will disproportionately contribute to overall virtual schooling achievement.

Papers such as *Can Virtual School Thrive in the Real World* (Wang et al, 2014) focus heavily on analyzing virtual schools with high asynchronous content provision and relatively lower peer effects and tend to find that virtual schools underperform traditional schools. Florida Virtual and many of the cyber-charter schools run by education management companies such as K-12 Inc tend to provide students access to digital content that can be consumed at their own pace. A high proportion of education hours in a typical week are self-service by the student with very little compulsory group class attendance and as a result a substantially reduced opportunity for peer interaction. One interesting area for analysis would be a results decomposition of these two broad categories of virtual schooling to see if there is indeed a bifurcation of results and schools that foster stronger peer-interaction drive academic achievement and help compensate for the weaker results of the schools that allow students a more fully self-paced, independent option.

The Causal Impact of Virtual Schooling on a Given Type of Learner Remains Unclear

It is unclear if virtual schools systematically result in any change on student achievement holding student quality and environment constant. While some research

(Wang et al, 2014) argues that virtual schools underperform brick-and-mortar schools, it is difficult to disentangle the causal nature of this relationship and compare fairly across student achievement at different institutions given the lack of randomized control trial experimental design and subsequent difference in the cohorts opting into the different learning pathways. A similar debate exists in the online for-profit university sector where many universities that are online have low graduation rates but also service second-chance learners who historically haven't performed well in traditional school environments and often have to balance working while studying. Is it fair to argue that these universities underperform when the baseline abilities of students are not taken into account and as a result academic outcomes are confounded relative to traditional universities? It would appear possible from the systematic review that students who are self-motivated, with strong parent/guardian involvement and a desire for academic performance could potentially outperform in virtual schooling environments while simultaneously students who are not motivated, have limited peer interactions or weak parental involvement would underperform in virtual schools or any school with reduced monitoring and active management for similar reasons.

The strong demand for virtual schooling, pre-COVID, in the United States, even at substantial cost to the parent, would suggest there is a market-based need for this pathway. Post-COVID, the need for virtual schooling is apparent to most school districts as an insurance against other forms of pandemics. Currently, billions of dollars annually are spent by governments and individuals on virtual schools. Given the heavy expenditure on virtual schooling, it is likely that for some parents, virtual schooling appears to be noticeably superior. However, the amount of spending on traditional brick-and-mortar private schools is substantially higher than total spending on virtual schools, so it may be the case that neither option is strictly superior or inferior in general but that brick-and-mortar is still heavily preferred for the significant majority of parents. It appears that virtual learning, at least,

offers a compelling set of characteristics for certain types of learners. If virtual schools were strictly inferior, there would be no or limited demand for private pay virtual schools. While one could argue that existing demand is present only because of the experimental, recent nature of these schools, the older ones have been in operation for more than a decade which is generally longer than the commercial life cycle of failed experiments (Sequoia, 2019).

The Importance of Mentorship

A consistent theme driving virtual schooling success is the importance of mentorship. This links to the previously discussed point about the need for parental involvement. Campbell et al (1997) highlights that strong mentorship is associated with higher GPA, higher satisfaction scores and higher completion rates. Additionally, in the online world, Drysdale et al (2014) highlights that virtual communication allows mentors to have a potentially closer relationship with students through electronic communication, texting and other forms of informal support and encouragement which may help to foster a stronger connection between mentor and mentee. This also aligns with the research suggesting an important driver of retention of teachers in virtual schooling is the quality of the student-teacher relationship. The mentor in this context would be defined most naturally as a teacher that the student engages with regularly but could also take the form of a tutor (who provides supplementary help on top of traditional online group classes with peers), or a more senior or academically advanced student.

2.9 Conclusion

To summarize, our systematic review into what drives student outcomes and student satisfaction in virtual schooling suggests (1) parental engagement (2) self-motivation in students (3) strong relationships with mentors and (4) peer effects play an important role.

Fruitful opportunities for further analysis exist in the delivery and production of effective online curriculum and in how to leverage technology to provide students with more feedback than traditional schooling in order to drive stronger academic attainment. The importance of the mentor is a powerful driver behind my logic for deploying psychometric matching to optimize the mentor-mentee match and drive the virtuous cycle of high teacher retention rates, high course completion rates and strong academic achievement by systematically increasing the probability a student forms a strong bond.

Peer-effects may be usable in the context of large-scale online schooling to enhance academic achievement by streaming across large populations of students. In many schools, there is not a large enough set of permutations of student classes and teachers to enable significant filtering by academic achievement. In schools of particular sizes, it may be easier to create classes with students of comparable ability which may induce enhanced academic achievement as peers of relatively comparable level mutually accelerate one another through healthy competition.

The findings around self-motivation of students suggest that online schooling may be a source of education inequity. It is possible that online schooling, given the removal of some of the constraints of physical schools that can take up time (such as assemblies, commuting and various types of non-academic peer interactions), may lead to enhanced academic outcomes in focused students who are liberated to focus more purely on their performance while at the same time, causing weaker academic students to underperform given the absence of some of the typical control

mechanisms (more easy access to web browsers and less ability for teachers to monitor students between classes as examples).

Additionally, the dynamics of providing student feedback in the online world offer potentially compelling opportunities for enhancing student achievement. Student Perception of Teacher Feedback and the Relationship to Learner Satisfaction in a High School Online Course (Lemmon et al, 2014) highlighted that the student's perception of the frequency and quality of feedback is the most important driver of academic achievement. In the online environment, students can be given faster feedback from teachers (such as a "digital emoji") or classmates without disrupting the class or creating unnecessary attention. This can help fuel the student's excitement and engagement without reducing the available class time for content delivery.

While virtual schools are still a relatively nascent type of education, this systematic review provides a survey of the key drivers of student outcomes and satisfaction and would suggest that synchronous delivery models in general may be more beneficial for students than the asynchronous, video-lecture based models that defined early endeavors in this space.

3

Systematic Literature Review: Use of Quantitative Matching in K-12 Education

3.1 Introduction

As the discussion around "personalised learning" grows, there is an opportunity to evaluate cost-effective methodologies to boost student satisfaction or student outcomes. Matching, which I define as an explicit attempt to curate individuals together to enhance an outcome of some kind, has relevance. If an online learning network, for example, has thousands of students and thousands of tutors and through the use of a cost-effective algorithm can better match these individuals together, student outcomes could potentially be improved at limited additional cost. The need to focus on academic outcomes is straightforward. The most

effective interventions are likely to actually raise outcome achievement. Student satisfaction is also important because making education more enjoyable for students is a worthwhile goal in itself. Low student satisfaction levels may lead to churn of students or a lack of engagement with course material. While no definitive link has been established in the academic literature between student outcomes and student satisfaction, I will consider both, as improvements in either may be useful.

This chapter looks at the existing "matching" literature in the context of K-12 education to see what areas may exist for optimization. "Matching" is of growing relevance when one considers online learning or blended learning (in which a student takes classes across both a brick-and-mortar campus and an online environment). The available set of potential matches a given student has is constrained to the number of relevant teachers in a particular school in a brick-and-mortar campus. In the case of online delivery, the pool of available teachers or tutors becomes a much broader supply which could draw from multiple countries and timezones. This significant expansion in available supply makes the relevance of "matching" more important. Previously, even if an effective matching technique was discovered, the number of choices the student had was limited. If the adoption of online learning continues to grow, accelerated by COVID-19, the relevance of large-scale optimization of student-teacher matches becomes increasingly important to understand.

Matching has become a topic of interest for many economists in recent years. Harvard Business School Professor, Scott Kominers, has used matching algorithms to help optimize a wide variety of problem spaces ranging from the Boston public school matching system (in which prospective high school students get matched to public high schools) to the allocation of fire-fighters. Matching, typically, is most effective when there is a large sample size of participants on both sides of a market to be matched and there is readily available information that can be used

to help filter the participants. Matching has relevance to the job market (matching prospective employees to companies), developing friendships (matching potential friends to each other potentially in a new city or just in general), online retailing (Amazon marketplace matching low-cost suppliers to price-sensitive consumers), lobbying (helping companies find lobbyists with relevant domain expertise aligned to the same political ideology), dating or even tutoring. In this relatively new field of matching, Alvin Roth won a Nobel prize for his theory of stable allocations and the practice of market design in kidney matching. While his analytical tools specifically were applied to kidney matching in many of his papers, the methodology was cross-applicable to a wide variety of fields. While I focus on the narrow area of education matching in my systematic review, further research could evaluate the broader matching landscape for relevant insights to optimal student-teacher matching as these insights may not necessarily arise from education specific research.

This systematic review follows my earlier systematic review on student outcomes and student satisfaction because it considers a much more narrow question: the use of education matching to drive student outcomes and student satisfaction. Given that education matching is a relatively nascent field of research, it was important to first consider broader drivers to be able to understand the context for why some of the matching interventions may or may not work. It also provided me with the opportunity to uncover a well-scoped niche within the matching literature that had limited contribution but appear promising: psychometric matching.

The use of algorithmic matching between students and teachers in brick-and-mortar schools was not the most natural problem space to explore because of the limited pool of participants on each side of the marketplace in a given school. Additionally, information on students and teachers is not always readily available to use as factors to determine matches. Online learning increases the scale of both sides of the marketplace and the available data on both students and teachers

making the problem space increasingly relevant for economists, educators and public policy makers alike.

Looking at the existing literature, three major types of matching have been considered. Firstly, gender matching, in which students and teachers are matched based on their explicit gender (Cho et al, 2012, Mwanza et al, 2017, Marsh et al, 2008). This has so far largely considered only male and female gender classifications to date. Secondly, ethnicity matching, in which students and teachers are matched based on a consideration of the participants' ethnicity (Banerjee et al, 2017, Eddy et al, 2011, Klein et al, 2001). Thirdly, personality matching, in which students and teachers are matched based on some psychometric or personality-based factor (Packer et al, 1998, Caine at al, 2017, Karch et al, 2007). In the existing literature, ethnicity matching doesn't appear to drive any impact on student satisfaction or student outcomes. Gender matching, within some of the studies, drove statistically significant positive improvements in student outcomes in Sweden and Greece (Cho et al, 2012). Personality matching has been shown to improve student satisfaction (Packer et al, 1998), but no work has been performed on student outcomes.

I am primarily interested in algorithmic matching in which the matching procedure can be conducted automatically without the need for human intervention. This is more useful for large-scale matching procedures and results in limited to no marginal cost for users of the algorithm. The literature has many examples of curated matching procedures involving a manual process on a small-scale but these aren't replicable for a large-scale school system or online education platform.

I am not interested in matching factors such as availability of a given teacher or current class size. These types of matching procedures can efficiently allocate students to under-capacity teachers which is useful when it is relevant but they aren't systematically optimizing for an increase in student outcomes per se, rather for capacity management. I am focused on matching algorithms that enhance

student outcomes or student satisfaction assuming there are no easy wins to be obtained from matching to under-capacity teachers.

I also focus on paid mentor-mentee relationships. In the school system, all of the educators are paid to be delivering sessions. Therefore, the most critical mechanism to understand is how to drive effective matching between students and a mentor/teacher who is paid. When the educators are paid to be delivering sessions, factors such as churn rates of mentors, absenteeism issues, level of preparation are less likely to be an issue because people are compensated for their time as opposed to acting exclusively altruistically. Matching algorithms that use volunteers are likely to be less relevant given the big variance in incentives between a volunteer and a paid educator and their reasons for conducting their work. It may also be more difficult to scale interventions that rely on volunteers. Finally, most individuals a student engages with along their educational journey are paid so this is a more pressing segment to consider.

I focus on the K-12 context. A number of inherent differences exist between university students and high school students. University students self-select to continue to pursue higher education and do so at substantial economic costs and non-monetary costs (effort, time). While high school is compulsory in many geographies, it is not always compulsory so some self-selection can occur at this level. University is generally optional in virtually all countries unlike high school. Additionally, the delivery model of many universities often revolves around large-scale lectures with less interpersonal interaction between faculty and students. Many of the largest private education organizations exist to support high school students and class size ratios tend to be smaller at high schools than universities (Department of Industry, Innovation, Climate Change, Scientific Research, and Tertiary Education, 2013). While some valuable insights can be garnered from the university matching literature, I focus explicitly on matching research from the K-12 space to avoid

conflating two fairly distinct pools of learners.

Finally, I consider the problem space of one-on-one matching. I first need to show one-on-one matching has the potential to improve student outcomes before moving to more complication group matching. Group classes make the social interactions between students and teachers substantially more complicated than a one-one context. In a student-teacher relationship there is only one relationship to examine. However, in a class of n students and one teacher the number of relationships is C(n+1,r)=(n+1)!/(r!(n+1–r)!). In a simple case of three students, Anna (A), Bob (B), Charlie (C) and a teacher Derek (D), there are six combinations including AB, AC, AD, BC, BD, CD. Additionally, there are second-order effects. For example, the relationship between Anna and Charlie in a one-on-one environment may be different to the relationship of Anna and Charlie in the social presence of Bob. This is understandably a very complicated problem to optimise for empirically and would require a substantially larger data-set. A matching algorithm that can optimize group class student-teacher allocations would be more useful to the current K-12 environment than a one-one matching algorithm, but is far more challenging to design.

This chapter will outline (1) the question addressed in the systematic review of K-12 matching, (2) the methodology used in the systematic review, (3) the summary tables of the systematic review, (4) key findings from the research papers assessed and (5) a discussion of the key themes emerging from the matching literature.

3.2 Systematic Review Logic

The following search terms and inclusion/exclusion criteria will be used for our systematic review. A systematic review is a rigorous compilation of evidence from all primary research studies within a defined set. We have chosen to do a systematic review given the scarcity of the literature, the reduced human bias in the process, increased repeatability because of the explicit methodology and a systematic presentation of synthesis of the characteristics and findings of the included studies.

The following search terms and inclusion/exclusion criteria will be used for our systematic review. This search was performed on November 21st 2018.

("Student" OR "Students" OR "Pupil" OR "Child") AND ("Teacher" OR "Mentor" OR "Tutor" OR "Instructor") AND ("Matching" OR "Pairing" OR "Selection")

("Mentee") AND ("Mentor") AND ("Matching" OR "Pairing" OR "Selection")

Questions addressed in systematic review:

Has any useful methodology been established for matching students and teachers to enhance student satisfaction or outcomes?

Method for Inclusion/Exclusion

Our inclusion criteria for student-teacher matching include:

1. Availability as a journal article between 1995 and 2018 in English,
2. Consider students engaging in K-12 schooling in either a brick-and-mortar school or an online school (between ages five to eighteen) across any type of study design,
3. Employ a matching algorithm with some quantitative methodology other than availability or class size to decide which students work with which teachers,

4. Consider teachers that are either part-time or full-time and either certified teachers in their state or country or non-certified teachers such as tutors.

Our exclusion criteria for student-teacher matching include:

5. Availability as a journal article in 1995 or prior to this year,
6. Availability in a non-English journal,
7. Consider students engaging in online university education,
8. Employ a matching algorithm with some quantitative methodology about availability or class size to decide which students work with which teachers,
9. Employ a matching algorithm that is qualitative.

Our inclusion criteria for mentor-mentee matching include:

1. Availability as a journal article between 1995 and 2018 in English,
2. Consider mentors in the context of an institution such as a government, business, school or leadership program,
3. Employ a matching algorithm with some quantitative methodology other than availability to decide which mentors work with which mentees,
4. Considers only mentor-mentee relationships in which the mentor is paid as part of their affiliation to the institution i.e. an executive who receives a salary mentoring a younger person (although not necessarily formally compensated for the specific act of mentoring).

Our exclusion criteria for mentor-mentee matching include:

5. Availability as a journal article in 1995 or prior to this year,
6. Availability in a non-English journal,
7. Considers mentors in the context of an institution such as a government, business, school or leadership program,

8. Employ a matching algorithm with some quantitative methodology based on availability to decide which mentors work with which mentees,
9. Employs a qualitative methodology for the matching algorithm to decide which mentors work with which mentors,
10. Considers mentor-mentee relationships in which the mentor is not compensated monetarily.

I conducted electronic searches in the following databases:

1. JSTOR
2. ProQuest Education
3. PsychInfo
4. ERIC
5. Ovid
6. EBSCOHost
7. Google Scholar

In total 14 electronic searches were conducted (seven search engines multiplied by two search criteria equalling 14) and the information was stored in "Covidence", a systematic review management software.

The Cambridge Quality Score Assessment Framework was used to assess the quality of the various articles in order to appropriately weight findings. This framework consists of three criteria: correlate score (measured by assessments of sampling, response rates, sample size, measures of correlate and measures of outcome), risk factor score (based on the type of data collected) and causal risk factor score (based of variation of risk factor and the analysis of change). Murray et al (2009) designed the Cambridge Quality Score Assessment Framework to aid in "identifying high-quality studies of correlates, risk factors and causal risk factors for systematic reviews and meta-analyses". In developing this framework, correlation

was defined as variables that have been shown to have an association with one another, risk factors was defined as variables that have a predictive relationship with the outcome because of clear temporal ordering and finally, causal risk factors are defined as risk factors that are variable which cause a shift in the risk for the outcome when they vary (Kraemer et al, 2005).

In the first component of the Cambridge Quality Score Assessment Framework, the checklist testing for correlates has five considerations which are binary in nature. This component of the framework is designed to the methodology by which the sample was collected, the response rates and the retention rates of people in the trial, the total sample size achieved and the methodology for assessing how the correlate and outcome were assessed.

In the second component of the Cambridge Quality Score Assessment Framework, the checklist is testing whether a given variable can be defined as a risk factor. In order to assess this, reviewers are asked to assess how the ordering of data in the study occurred. Cross-sectional data, which is the least rigorous, is given a value of one. Time-ordered retrospective data is given a value of two. Prospective longitudinal data in which a risk factor is given scores a three.

Of the three components of the Cambridge Quality Score Assessment Framework, the largest contributor to overall score and subsequently the most significant element in the assessment criteria is the causal risk factors. This part of the framework is testing for common issues with causality in non-randomized studies. The first consideration is whether within-individual changes in a given outcome variable are also associated with within-individual changes in the specified risk factor. The second consideration is whether the study has controlled for alternative mechanisms to explain the findings. In order to score a full 7/7 in the assessment framework, one must conduct a randomized control trial with a specific targeted risk factor. In the absence of full randomization, a 6/7 can be scored if variation in the risk-factor

is related to within-individual changes in an outcome variable while controlling for all relevant confounding variables.

Alternative assessment frameworks including the Maryland Scientific Methods Scale (Farrington, 2003) was considered, but ultimately in considering the set of research papers, the Cambridge Quality Score Assessment Framework was the most compelling choice.

3.3 Summary Tables of Systematic Review

The table below summarizes the various findings of the systematic review.

Table 1 - Reference: Each research article is matched with an ID to assist with navigation across the tables

	Research Title	Description of Participant Population	Intervention
1	Student-Teacher Ethno-Racial Matching and Reading Ability Group Placement in Early Grades	Nationally Representative Sample of Kindergarten and First Grade (17% African American, 20% Latina, 62% Whites)	Teacher-Student Ethno-Racial Match
2	Speed Dating for Mentors: A Novel Approach to Mentor/Mentee Pairing in Surgical Residency	Resident Doctors Rotating Across Four Hospitals in the North-East	Senior residents met with junior residents in 90 second intervals and then ranked top three matches
3	Effect of Teacher-Student Gender Matching: Evidence from OECD Countries	Students and Teachers across 15 OECD Countries	Teacher-Student Gender Matching
4	Teacher-Student Matching and the Assessment of Teacher Effectiveness	Fifth grade students in North Carolina	Teacher-Student Matching
5	Ethnic Matching, School Placement and Mathematics Achievement of African American Students from Kindergarten through Fifth Grade	Kindergarten to Fifth Grade African American students across the USA	Student-Teacher Ethnic Matching
6	Finding optimal mentor-mentee matches: a case study in applied two sided matching	Theoretical Simulation	Preference Matching
7	Meet-n-Greet: a mentor-mentee matching approach for increasing the prevalence of naturally self-selected mentoring partners in program based matches	n/a	n/a
8	Does Matching Student and Teacher Racial/Ethnic Groups Improve Math Scores?	Fourth grade students and teachers in California	Racial/Ethnic Group Matching
9	A Multi-Level Perspective on Gender in Classroom Motivation & Climate: Potential Benefits of Male Teachers with Boys	Year 8 and Year 10 high school students in Math, English and Science classes	Gender Matching

10	Estimating Causal Effects of Teacher-Child Relationships On Reading and Achievement on a high risk sample	Kindergarten children	Effect of High-Quality Teacher Student Relationship
11	Effect of Gender Concordance in Mentoring Relationships on Summer Research Experience Outcomes for Undergraduate Students	Building Undergraduate Infrastructure Leading to Diversity University Students	Gender Matching
12	Ethical Behaviors of Student Teachers' Mentors in Forced Same-Gender and Cross-Gender Matches	Student Teachers in the Second Term of Teaching Practice Mentoring Program	Gender Matching
13	Cognitive Style and Teacher-Student Compatibility	Psychology Student and Trainee Math Teachers	Cognitive Style Matching
14	Does Race-Matching Matter?	Urban High School Teachers and Students	Racial and Gender Matching
15	Individual Differences in Preferences for Matched-Ethnic Mentors Among High-Achieving Ethnically Diverse Adolescents in STEM	Ethnically diverse adolescents in STEM	Preference for Racial Matching

Table 2 - A List of Varying Studies, Outcomes, Settings, Participants, Control Groups, and Findings

	Comparator	Outcomes	Setting	Participants	Assumed Control Group Risk and Corresponding Intervention Group	Mean Difference or Standard Mean & Confidence Interval
1	Three-Level Hierarchical Linear Regression	Reading Ability Group Placement	School	23670 (Kindergarten Sample 11,260 and First-Grade Sample 12,410)	Kindergarten M: 0.68, SD: 0.52, Kindergarten Match M: -.02 SD: 0.03 First Grade Control M: 0.69, SD: 0.61, First Grade Match M: -0.07, SD: 0.05	Kindergarten Mean Difference: 0.7, First Grade Mean Difference: 0.76
2	Fisher's Exact Test Comparison	Cross-Sectional Likert Survey with Univariate Analysis of Satisfaction	Academic General Surgery Residency Program in the Northeast of the United States	29 Postgraduate Year 1-2 JR and 28 Year 3-5 Senior Residents	Control Group Satisfaction: 12%, Speed Dating Satisfaction: 85%, Exact chi-squared test P=0.0012	Difference in Satisfaction: 83%
3	First Difference Model	Test Scores	Schools (~ 3/4 of students in 8th grade and 1/4 in 7th grade)	202,644 lower secondary students from 15 OECD countries	Control Group Mean: 0, SD: 1; Hungary Mean: 0.02, SD: 0.01, Spain Mean: 0.12 SD: 0.05, Greece Mean: 0.04 SD: 0.01, Sweden Mean: 0.02 SD: 0.01	Difference in Mean: 0.01
4	Linear Model (Estimation of Education Production Function)	Standardized Test Score	Schools	60,971 Fifth Grade students in North Carolina and 3,223 Fifth Grade teachers	Male Teacher Mean Decrease for Reading: -0.032 SD: 0.018, Hispanic Teacher Mean Decrease for Reading: -0.113 SD: 0.036	1 SD increase in teacher licensure score increases predicted student achievement in math by 1-2% of a SD
5	Two-Level Growth Model	Mathematics Item Response Theory Scores (IRT)	School	1200 African American students	Control Group with 0 African American teachers had a mean of 2.14 compared to Group with at least 1 African American teacher had a mean of 3.01	Difference in Mean 0.87 but not statistically significant as CI [1.99 3.73]
6	Simulation	Average Welfare Across Pairs	Simulation	100 independent repetitions	n/a	No Guarantee of match has complexity (L) where L is sum of length of preferences. GATA guarantee is complexity (pL) where p is population factor

7	n/a	n/a	n/a	n/a	n/a	n/a
8	Individual Level Regression Equations	1999 Stanford 9 Math Scores	Schools	136 California elementary schools, 281 4th grade teachers	Control Group Coefficient: 0, White Teacher & Black Student: 0.124	Difference: 0.124 at the 5% significance levels
9	Cross-Classified Multi-Level Model with Five Levels	Motivation and Engagement Scale	Schools	964 high school students from five co-educational government schools	Control Group Coefficient: 0, Coefficient of Male Teacher for Boy Student: -0.02	No Difference at the 5% significance level
10	Multi-Level Propensity Score Matching Approach	Student Math and Reading Achievement	22 Elementary Schools in 3 inner-city districts	324 children and parents and 60 teachers	Effect of Teacher-Child Relationship on Math Achievement = 1.78, SE = 0.71, $p < 0.05$	Difference: 1.78 at 5% significance level
11	General Estimating Equations, ANOVA	Student Gains (A combination of three constructs on a 1-5 point Likert Scale of personal gains, gains in knowledge and skills and thinking and working like a scientist)	Undergraduate College	109 Built Undergrad Students	"Thinking + Working Like a Scientist" 0.136 p=0.024	Difference: 0.136 at the 5% significance level
12	Cross-Sectional Study	Prevalence of Ethical/Unethical Behavior	Teacher's College	616 Student Teachers	Control: Mean: 2.6 SD: 0.59 v Same: Mean: 2.7 SD: 0.57	0.1 not significant at the 5% significance level
13	Experimental Design	Objective Test Performance	Undergraduate College	54 final-year trainee math teachers and 58 first-year psychology students	Significant interaction between teacher's cognitive style and student's cognitive style $F(1,12) = 5.97$, $p<0.0$, $G(1) = 4.97$, $p <0.05$ indicating student's evaluation of teachers were influenced by teacher's cognitive style	0.24 at the 5% significance level
14	Hierarchical Linear Modelling	Student Report of Fairness, Trust in Teachers, Number of Reports of Exclusionary Discipline, Student Perception of Exclusionary Discipline	Urban High School	183 students and 19 teachers	Teacher's student-gender match led to lower likelihood of students receiving one or more disciplinary sanction (Beta = -1.45, p = 0.003)	Coefficient is -1.45 at the 5% significance level

15	Longitudinal Study	Self-Efficacy, identity and commitment to a science career	Residential science education program	265 students	F(4, 182) = 2.29, p=0.09, n_p^2 = 0.4. Stable high group had significantly more commitment than stable low	0.5 increase in identity as science student for those who viewed matching with same mentor as high. Not significant at 5% level

Table 3 - A List of Cambridge Quality Checklist Quality Scores and Quality Score Totals.

	Correlate Score	Risk Factor Score	Causal Risk Factor Score	Total
1	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Outcome: 1, Measure of Correlate: 0	2: Retrospective Data	3: Study without variation in risk factor	9
2	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 1	1: Cross-Sectional Data	1: Study without variation in risk factor with no analysis of change	4
3	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	5: Study with variation in risk factor and adequately balanced, No analysis of change	12
4	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Outcome: 1, Measure of Correlate: 0	2: Retrospective Data	3: Study without variation in the risk factor with analysis of change	9
5	Sample: 0, Response Rate: 1, Sample Size: 1, Measure of Correlate: 0, Measure of Outcome: 0	2: Retrospective Data	3: Study without variation in the risk factor with analysis of change	6
6	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	0: No Data	0: Theoretical study	0
7	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate: 0, Measure of Outcome: 0	0: No Data	0: Theoretical study	0
8	Sample: 1, Response Rate: 0, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	5: Study with variation in risk factor and adequately balanced, No analysis of change	10
9	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Outcome: 0, Measure of Correlate: 0	1: Longitudinal Data	4: Study with variation in risk factor and adequately balanced, No analysis of change	8
10	Sample: 0, Response Rate: 0, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	6: Study with variation in risk factor and adequately balanced with analysis	9
11	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	1: Cross-Sectional Data	3: Study without variation in the risk factor with analysis of change	7
12	Sample: 0, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	1: Cross-Sectional Data	3: Study without variation in the risk factor with analysis of change	8

13	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	3: Study without variation in the risk factor with analysis of change	8
14	Sample: 0, Response Rate: 1, Sample Size: 0, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	3: Study without variation in the risk factor with analysis of change	8
15	Sample: 1, Response Rate: 1, Sample Size: 1, Measure of Correlate: 1, Measure of Outcome: 1	2: Retrospective Data	6: Study with variation in risk factor and adequately balanced with analysis	13

3.4 Results

In total, there were fifteen research papers that satisfied the search terms and inclusion/exclusion constraints. Three of the research papers focus on gender matching, in which the effect of cross-gender and same-gender matches are examined. Five of the research papers focus on racial matching, in which the effect of mentor and mentees being of the same race and different races is examined. Two of the research papers leverage experimental social interactions followed by two-sided preference rankings of mentors. One of the research papers focuses on matching based on the academic ability of a student and academic experience of the teacher. One of the research papers examines various theoretical algorithms for matching and compares their efficiency in a simulated environment. Two of the research papers examine the perceived importance of the relationship as a factor driving success of the match. One of the papers examines psychometric matching (or cognitive style matching).

The research articles average scores of 7.40 with a standard deviation of 3.69. It is important to note the absence of prospective data and the absence of any randomised control trials.

In total between all search results in all journals across the two systematic reviews conducted, 699 articles were identified through database searching. 480 articles remained after duplication removal. 342 articles were then removed during title and abstract screening. Of the remaining 138 articles, 75 articles were removed during full text screening and 63 articles made it to extraction. Of the 63 articles, 15 research papers related to algorithmic matching and 48 research papers related to student outcomes and student satisfaction in virtual schooling. (Chapter Two, Beaton, 2021)

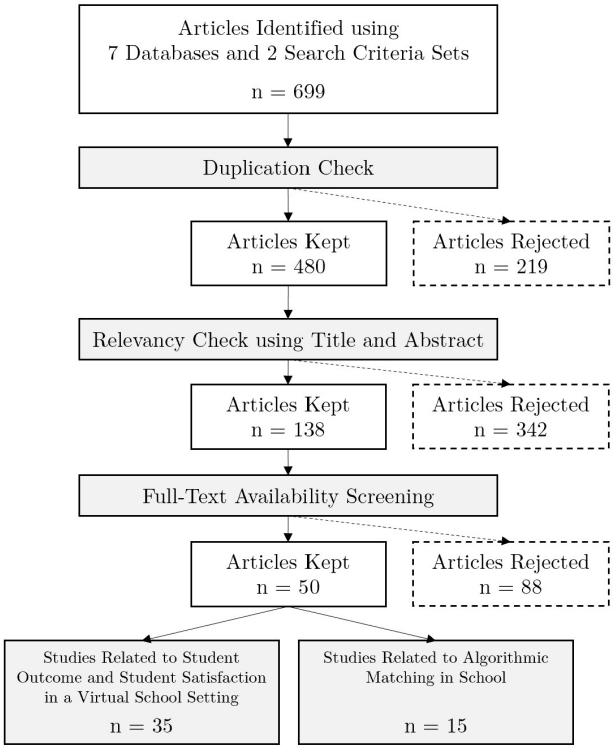


Figure 3.1: Article Search Process

3.5 Key Findings of Research Papers In the Systematic Review

Ethnicity/Race Matching

Student-Teacher Ethno-Racial Matching and Reading Ability Group Placement in Early Grades (Banerjee et al, 2017) found that same race matching has no effect on kindergarten students' reading class placement levels but a positive effect on first grade students' reading class placement levels. In other words, students who were paired with a mentor of the same ethnicity had a statistically significantly stronger likelihood of being placed in a higher reading class.

Ethnic Matching, School Placement and Mathematics Achievement of African American: Students from Kindergarten through Fifth Grade (Eddy et al, 2011) assessed the impact of ethnic matching on a measure of mathematical aptitude

and found no statistically significant relationship between ethnicity of teachers and academic success of minority teachers. This analysis had a relatively low score on the Cambridge Quality Assessment largely because of a convenience sample with limited analysis of change of the risk factor.

Does Matching Student and Teacher Racial/Ethnic Groups Improve Math Scores (Klein et al, 2001) used individual level regression equations to evaluate the impact of ethnic matching and found no statistically significant relationship. After adjusting for various student and teacher characteristics, only teaching experience showed a statistically significant correlation with test scores. This paper was one of the most technically robust econometric papers and used a substantial number of control variables. The data set was relatively small so there were concerns that this was a convenience sample as opposed to a randomised sample.

Does Race-Matching Matter (Roberts Young et Al, 2018) did not find any statistically significant effects for race-matching but also considered gender matching. In the paper, it was observed that teacher-student gender matching led to a statistically significant decline in receiving one or more disciplinary actions (Beta: -1.45, $p = 0.003$). No significant relationship between teacher-student gender match and teacher trust or fairness. Race matching did not predict trust or fairness or likelihood of disciplinary event.

Individual Differences in Preferences for Matched-Ethnic Mentors Among High-Achieving Ethnically Diverse Adolescents in STEM (Goza et al, 2012) did not find any statistically significant effects for race-matching. The methodology considered independent variables including measures of self-efficacy, identity, and a commitment to a career in science. One interesting finding of the paper was that two-thirds of students who regarded ethnic-matching as important continued to share a belief in the importance of ethnic-matching after being matched. One-third, despite initially holding the view that ethnic-matching was important, switched to hold

the view that it didn't matter at all. The curious discrepancy between the lack of statistical significance in findings but the sustained belief of the majority of the cohort may be a result of unconscious bias or could suggest some unobserved factors are at hand that were not effectively tested for.

Gender Matching

Effect of Teacher Student Gender Matching: Evidence from OECD countries (Cho et al, 2012) performed a comprehensive analysis of 202,644 lower secondary school students from OECD countries and evaluated the impact of same-gender or cross-gender matching on test scores. They found statistically significant effects at the 5% confidence level of matching girls with female teachers in Greece and Sweden on higher academic performance. They also found statistically significant effects at the 5% confidence level for matching male students with male teachers in Hungary and Spain on higher academic performance. This study performed well in the quality score assessment with a score of 12.

Ethical Behaviors of Student Teachers' Mentors in Forced Same-Gender and Cross-Gender Matches (Mwanza et al, 2017) assessed whether or not gender-matching led to shifts in the prevalence of ethical or unethical behavior. While some differences were found, none of these differences were statistically significant. This study was distinctive in it being the only paper in the systematic review that analysed ethical behavior. Judging ethical behavior is an ambiguous concept and various conditions were considered including (1) hardworking in mentoring, (2) using power and authority to expose me a risky situation, (3) seeks my permission for action on my behalf, (4) confidential, (5) fair, (6) does not conflict his interest, (7) caring. Although this analysis didn't show any statistically significant differences, there was considerable variance between same-gender and cross-gender pairs and it would be useful to explore this hypothesis through a variety of other methods.

A Multi-Level Perspective on Gender in Classroom Motivation & Climate: Potential Benefits of Male Teachers with Boys (Marsh et al, 2008) found no statistical effect of gender-matching on academic performance at the 5% significance level. One interesting finding was that the class environment was a more significant factor in driving academic performance than the teacher in this study. This study had a reasonably large survey size of 964 high school students and students straddled math, science and English classes.

Personality or Psychometric Matching

Cognitive Style and Teacher-Student Compatibility (Packer et al, 1998) had one of the stronger statistical relationships observed in the systematic review. Significant interaction was observed between teacher's cognitive style and student's cognitive style indicating student's evaluation of teachers were influenced by teacher's cognitive style which was significant at the 5% significance level. This was the only paper in the sample that looked at personality as a driving factor of the mentor-mentee interaction. It had a relatively small sample size of 54 final-year trainee math teachers and 58 students but was still significant at this level.

Speed-Dating for Mentors: A Novel Approach to Mentor/Mentee Pairing in Surgical Residency (Caine et al, 2017) recommended a strategy where mentors and mentees met socially for a series of rapid-fire interactions then were asked to rank a list of the top three individuals (in no preference order) they were most excited to work with. Following this, individuals were matched to maximize the number of successful matches (based on two-sided interest). The cross-sectional analysis showed that speed-dating led to substantially higher satisfaction. A serious concern with this analysis was the small sample size which resulted in a low quality score assessment of 4.

Meet-n-Greet: a mentor-mentee matching approach for increasing the prevalence

of naturally self-selected mentoring partners in program (Karch et al, 2007) based matches did not offer any compelling quantitative arguments but used a case-study to argue for the importance of initial mixing of potential mentors and mentees. Following a short trial interaction, the paper argued that the chance of a positive mentor-mentee experience is substantially increased. This paper received a very low quality score given its absence of sample size or credible statistical analysis.

Teacher Qualification Quality

Teacher-Student Matching and the Assessment of Teacher Effectiveness (Clotfelter et al, 2006) assessed standardised test scores and found a positive correlation between strength of teacher qualification and teacher achievement. This was driven by sorting of teachers across schools and to a lesser extent with schools. A one standard deviation increase in teacher licensure score predicted an increase in student achievement by 1-2% of a standard deviation.

Other Matching Considerations

Finding Optimal Mentor-Mentee Matches: A Case Study in Applied Two Sided Matching (Haas et al, 2018) was a purely theoretical analysis that considered the computational complexity of various matching algorithms for mentor-mentees. The paper made a compelling argument as to why the processing time to process a series of matches (without a guarantee of matching all pairs) in which each candidate has a list of preferences. L is a function defined as the sum of the length of the preferences. Out of the various algorithms tested, the GATA-Mixed Heuristic was the most efficient matching algorithm occuring in PL where P is a population factor. This can be best described as an optimised weighting of a variety of other algorithms and results in a guarantee of completed matching.

Estimating Causal Effects of Teacher-Child Relationships On Reading and

Achievement on a high risk sample (McMormick et al, 2013) was one of the higher quality publications because of its more comprehensive statistical analysis in assessing different strengths of teacher-child relationships. The paper found that for first grade, low-income students attending an urban school, high-quality teacher student relationships had a statistically significant positive impact on math achievement.

3.6 Discussion of Key Themes Emerging from the Systematic Review

Generally speaking, the literature for student-teacher matching is fairly sparse. The two most commonly appearing types of matches are race-based matching and gender-based matching. I will first summarize some of the key findings from the race-based matching literature.

In an American context, the achievement of African American students is a big concern for public schools and policymakers. While political initiatives have been passed, such as *The No Child Left Behind Act*, which forces schools to be responsible for the outcomes of all students. Despite substantial focus, many schools have struggled to enhance achievement of any underrepresented minority students systematically including African American students. As a result, there has been a big increase in qualitative and quantitative analysis into the potential for ethnic matching of underrepresented minorities to enhance student outcomes systematically. This would be achieved, for example, in assigning an African American teacher to African American students in an academic setting for a subject like mathematics or as a mentor on top of existing classes.

Analysis by Dee (2004) and Easton-Brooks, Lewis & Zhang (2010), showed that African American teachers had a positive effect on the reading outcomes of elementary public school African American students. Regardless of subject, researchers from Dee (2004) to Foster (1997) to Henry (1998) have demonstrated compelling connections between culture and knowledge in the context of learning and knowledge acquisition of underrepresented minorities. The most common theoretical rationale backing the findings of these studies is the discordance between the home culture and the school culture of students (Banks (1996), Milner (2007), Nieto (2000)). Many researchers have argued that teachers from minority backgrounds are

better able to bridge the discordance between home and school culture including Lewis (2006) and Shipp (1999). Research from Zimmerman (1998), Casteel (1998) and Ferguson (1995) found compelling evidence that teachers often evaluate the academic achievement of students of their own gender more favorably than students of differing ethnic groups. The most popular explanation for this is the impact of the perceived cultural similarities between teachers and students. This is not necessarily a good thing if teachers systematically mark-up the exam results of students from the same ethnic group because of subconscious bias but may be possibly useful if the results are because of enhanced cultural context being applied to the arguments being made, especially in humanities subjects in which interpretation plays a bigger role.

A crucial theoretical rationale for why underrepresented minorities perform better when matched with teachers of comparable backgrounds is that teachers have an enhanced perception of the student's ability and potential to learn. Dee (2005) argued that if a student is not matched by ethnicity, they are likely to be perceived by teachers as disruptive and lacking potential to learn. This effect intensified when considered in the context of lower socioeconomic achievement.

Ferguson (1998) found that Caucasian American teachers, across a comparable piece of academic work, perceived the academic quality of work produced by African American students to be lower than when assessed by an African American teacher. Roberts-Young (2018) found that when cultural background was not matched between student and teacher, there is an increase in the use of disciplinary force.

An additional complexity is that underrepresented minorities tend to be disproportionately found in schools, which have high concentrations of underrepresented minorities as opposed to being spread in proportion to their population representation across schools. Brown-Jeffy (2008) found that if at least half of the students in a school are ethnic minorities (defined in this case as African American or Hispanic), the achievement of all students is lower on average than when students attend a

school with a lower proportion of minority students.

Some of the more specific empirical findings that directly tackle ethnic matching for students and teachers include Dee (2004). He found that reading and math scores of African American elementary students were higher when matched with an African American teacher for a year. Clewell et al (2005) provided evidence of an increase of reading and mathematics scores of fourth grade African American students who received instruction from African American teachers. One of the most quantitatively robust studies I reviewed was Eddy and Easton-Brooks (2011) which found by studying the Early Childhood Longitudinal Kindergarten-Fifth Grade data set that if a student had at least one teacher who ethnically matched themselves between kindergarten and fifth-grade, there was a significant impact on mathematics achievement. This result, as expected, was more pronounced for underrepresented minorities than European students. European students generally have more community role models and public role models, which could explain why the effect is less pronounced being the majority of the population of American. Egalitea et al. (2015) found that African American and White students score higher in math and reading when assigned to teachers of the same racial background. This particular research didn't find conclusive results for ethnic-matching of Latino students. Social psychology research which attempts to explain these empirical findings varies but a common theme is that students' feeling of attachment and sense of belongingness to schools are enhanced by teachers from the same cultural background (Crosnoe et al., 2004, Johnson et al., 2001). More broadly, context and culture matters and that findings vary quite significantly from ethnic group to ethnic group.

Downey & Pribesh (2004) argued that because teachers spend a significant amount of time each day with students, racial biases in the evaluation of student performance may have meaningful adverse implications to the school experience of

minority students. A particularly rigorous paper, which did a meta-analysis of 32 research papers highlighted that teachers generally held the highest expectations for White and Asian-American students, followed by Latino students with the lowest expectation being African American students (Tenenbaum & Ruck, 2007). A fascinating dimension of this research is the consideration of what exactly constitutes racial bias. Three standards of bias are offered: 1) unconditional race neutrality is an environment in which teachers expect the same from all students without any consideration of performance or unobserved potential. This is practically very rare in the context of teaching, 2) conditional race neutrality is when teachers base expectations purely on a student's past performance. Teachers can consider race in forming their expectations of students, only if they accurately assess the historical performance of a given race and do not overestimate or underestimate and finally 3) race neutrality conditioning is when a teacher forms perceptions and expectations based on the proven historical performance or on latent potential, even if it varies by racial group. Under this third definition, a teacher is considered biased if they underestimate either of these factors systematically. Banerjee (2017) found that teacher assignment of students into ability groups (streaming) is biased racially, whereby Latina teachers systematically placed Latina students into higher ability groups than European teachers. In the same paper, it was also found that teachers' perception of students' learning attitude and behavior (which social psychology literature links to ethnic matching) has a positive correlation with streaming group placements. Considering the general evidence that racial inequality in education begins in the very beginning of kindergarten and early grades and then continues to grow, the systematic bias found in the placement of young children academically is problematic.

The literature on gender-based matching is another area of algorithmic matching which has received significant attention. Cho (2010) produced one of the most extensive summaries of the gender-matching literature by evaluating the impact

of gender on student-teacher outcomes across all OECD countries included in the research. This analysis, which employed extensive use of fixed effects, found that in eight out of fifteen OECD countries, same gender matching had no impact on student outcomes. However, in five out of fifteen OECD countries, teacher gender is positively correlated with student's test scores at the five percent significance level. A further two OECD countries had significance at the ten percent significance level. Much of the older literature on this topic argued that teachers of the same gender would be more effectively able to communicate with students and instill higher performance expectations. This was explored by Braun (1976) and Rosenthal & Jacobsen (1968). Another explanation offered was that teachers of the same gender as the student could be a more effective role model (Angrist et al, 1971). Some of the literature argues as far as saying that opposite gender-matches actually leads to detrimental results on student outcomes by intensifying negative gender stereotypes (Steele, 1997).

More recent empirical results are also contentious. Dee (2007) noted that if a teacher is matched with a student of the same gender, they are more likely to rank them higher on subjective measures of performance. Particularly for females, some studies have shown that gender matching leads to enhanced academic achievement and more appropriate course selection (Keil & Russo, 1998; Nixon & Robinson, 1999). Other studies such as Krieg (2005) have concluded no such effect.

Mentor gender is also argued to affect the process and outcome of mentoring (Kesser & Allen, 2010). Young et al (2005) made the argument that mentees in gender-matched relationships may experience greater comfort and more psychosocial support. The evidence on this is still fairly uncertain and a well-regarded study into publication rates by Ugrin et al (2008) found gender-mismatched faculty and students had higher publication rates than same gender matches for men and women. As noted above, much of the research is focused on outcomes for women and there

is little evidence to suggest men receive more psychosocial support from women over men or vice versa in mentoring relationships (Young et al, 2004).

After race-based matching and gender-based matching, the systematic review also highlighted a number of interesting additional findings. The most noticeable were findings around cognitive style and teacher compatibility. Over the last several years, there has been a growing embrace by the private sector of personalised learning (supported by initiatives such as the Chan Zuckerberg initiative, which seeks to provide socially-conscious capital to drive innovation in education-technology). While educators have focused for a long time on needing to understand the baseline achievements of a student and assessing their progression from there, many organizations, particularly in the private sector, have been slow to embrace this thinking. For years, companies have employed measurements of personality such as psychometrics to assess characteristics of potential employees beyond purely academic characteristics. There has been limited application of these techniques to education.

This paper argues convincingly that the student's objective learning outcomes, perceived evaluation of the ease of learning (an indirect measure of satisfaction), teacher's ability to communicate with the student easily and were more accurate in predicting the scores of students through cognitive-style matching. The experimental design leverages a 2 x 2 x 2 x 2 design that intentionally focused on testing extreme cognitive characteristics. As a result, the findings from this specific paper can only be generally applied to extreme personality characteristics. Furthermore, the relatively small sample size of the analysis which considered only 54 final trainee math teachers and 58 first-year psychology students would make it difficult to easily evaluate more marginal personality characteristics. This suggests the need for a larger-scale assessment of the impact of cognitive matching on student outcomes, student satisfaction and other measures to see the external validity of these results.

An interesting suggested direction from the paper was to evaluate the impact of mismatched cognitive styles between a student and a teacher of reducing the efficacy of communication styles. Finally, the paper suggests analysis is conducted into teacher-class matching rather than simply teacher-student matching. It is possible that a teacher may meaningfully adjust their style of teaching in a one-on-one capacity and a group capacity because it is no longer practical to personalise instruction to the needs of a single student if faced with a large class of students with varying needs and ability. A further analysis that may be useful would be to evaluate the consistency of the teacher's teaching style with differing numbers of students and also students of varying characteristics. I would hypothesize that more experienced teachers would generally make larger adjustments when confronted with different environments.

Two research papers addressed the idea of preference matching where students and teachers (or mentors and mentees) are exposed to each other and given the opportunity to submit a ranking. Speed Dating for Mentors (Caine et al, 2017), despite a very small sample size, found highly statistically significant results which suggested that satisfaction ratings could be meaningfully improved by giving people preference matching. While one could hypothesize that giving people an option to share their opinion then implementing the opinion is likely to lead to general satisfaction increases without necessarily generating a corresponding increase in outcomes, it is also possible that people can make informed judgements about the types of individuals that are socially compatible and engaging for them as these predictions are informative. While any individual factor such as gender may not be significant, it may be that matching along a spectrum of dimensions (i.e. similar cultural background, socioeconomic background and career interest) may generate a spike in engagement from both participants and drive a higher frequency of meetings and stronger two-sided accountability. In both these papers, the participants were aware they were matched and this may also be responsible for

driving a placebo effect of sorts in which they feel the person they are interacting with is supposed to be a good fit for them.

Building on the findings of cognitive style matching, the preference matching literature supports the idea that it is possible to systematically improve at least satisfaction and possibly outcomes by considering interpersonal dynamics. The question to ask is what is it that mentors and mentees see in one another that can be deciphered in a brief meeting and can this be systematically pre-identified before any meeting takes place? The importance of pre-identification is two-fold. Firstly, by requiring pre-identification it necessitates a more rigorous academic understanding of what drives the "magic" of successful mentor-mentee chemistry. Secondly, meeting mentors is only practical when the pool of potential mentors is fairly small. Finding optimal mentor-mentee matches (Haas et al, 2018) highlights the complexity of this second factor by theoretically modelling the processing time required for various types of matching algorithms. Considering a simple thought experiment, if there are a thousand potential mentors and a thousand potential mentees and each pair needs to meet each other at least once, even if all meetings can take place concurrently across the population of students, it will take 1000*(meeting time) to complete the initial assortment of preferences. Beyond the enormous time required to complete all these meetings, the sheer volume of interactions may reduce the ability for any one person to effectively create an ordered preference list. Kahneman's work on decision making (Kahneman et al, 1991) suggests that the mind has a limited capacity to remember preferences and rules of transitivity tend to break down above a certain number of options.

If one can be fairly confident that small, unstructured social interactions lasting a couple of minutes can convey enough information to drive meaningful improvements in matching, it would appear that algorithmic matching sorting individuals together based on an understanding of a certain set of factors such

as personality, interests, IQ and other factors could yield statistically significant outcomes if matches are conducted on the correct factors. Additionally, based on the work of Caine, it would appear that the unstructured social interactions generally result in mentors and mentees sorting towards people that resemble themselves as opposed to sorting towards people that vary significantly from themselves. This would seem to match with intuition from observed social interactions at university campuses with high proportions of international students in which students of similar ethnic groups tend to readily sort together through various student unions and informal social relationships.

In assessing the evidence on algorithmic matching for our systematic review, I have found variable relationships related to ethnicity based matching that don't clearly indicate an opportunity for algorithmic matching to systematically improve outcomes. While we find some tentative evidence that personality and gender matching may be interesting, this is based on small scale correlation studies. Randomized control trials will be useful to disentangle to test causal effects of gender and personality. While it is likely that a multivariate matching technique or a machine learning algorithm may offer the most promising results, the lack of large scale data collection makes for a difficult problem space to deploy such techniques. Additional tests of gender matching and cognitive style matching are likely to be fruitful areas for further exploration.

4

Qualitative Analysis

4.1 Summary

As part of my broader review of existing online schooling practices that drive student outcomes and student satisfaction, I have conducted qualitative interviews with a range of current online high school students engaging in different formats of online learning, either part-time or full-time.

This qualitative research process is critical to understand the lived experiences of online high school students and build my empathy as a researcher for the dynamics that affect these learners. Additionally, this contextual understanding is quite important for researchers seeking to understand the online schooling space as they may have a number of misconceptions about the environment with a research lens developed in a brick-and-mortar schooling context. Having conducted two

systematic reviews, this qualitative study is important to provide additional context and lived experience to add further detail to some of the earlier findings. They also provide an important human perspective so readers can build empathy with the types of students present in the quantitative studies later in the thesis.

Contrasting the explicit differences between a physical environment in which in-person social interactions are relatively frequent, people coalesce around a campus and have opportunities to bond with their classmates outside of an academic environment (through sports and social activities) to an online school, in which people may not ever physically meet their classmates is important to enhance researcher empathy. My work in this chapter seeks to help build understanding for why students choose online schooling environments in the first place and what are some of the common characteristics of this segment of learner. It also is informative to directly ask the student about the weaknesses and strengths of their environment and how they think it could be improved.

For my thesis, I conduct a randomized control trial and a cross-sectional analysis to optimize the student experience in an online context. It is important to make sure these interventions are reconciled against the real context and observations of students about their own environment. The interviews conducted in this chapter help to build context for why a given cognitive or personality based intervention may work otherwise it is hard to build intuition for the learners and what may enhance their outcomes through a purely abstract, empirical lens.

My primary research questions for this qualitative analysis were: 1. What factors lead a student to choose to pursue learning in an online high school? 2. What factors lead a student to be successful in an online high school? 3. What factors lead a student to struggle in an online high school?

After interviewing twenty-one students attending Stanford Online High School, the nineth highest ranked high school in the US, supplementary online learning

platforms, Florida Virtual and various New Zealand schools we find the following recurring themes in students' perceived requirements of success of virtual learning: (1) self-motivation, (2) desire for academic acceleration through extension coursework, (3) less stressful social and competitive environment (4) flexibility in schedule facilitating other activities. Similarly, I found recurring themes in students' perceived weaknesses of online schooling: (1) higher barriers for communication with teachers, (2) less spontaneous opportunities for social interaction, (3) limited access to school extracurriculars.

I will firstly describe the methods used for this research including the participants, procedures, analysis methods, ethical concerns as well the role of the researcher. I then discuss the key findings from my qualitative interviews.

My initial interviews were typically thirty minutes in length and respondents generally answered a range of pre-specified interviews. I then ran deeper interviews with an unstructured agenda designed to explore key themes from the initial batch of interviews more rigorously.

I then present my analysis using NVivo, a qualitative data analysis software. NVivo is typically used when rich analysis of text-based data is required. Specifically, it allows for analysis of unstructured data across text, audio, video and image across interviews, focus groups, surveys and journal articles (Kent State University, 2020).

4.2 Methodology

Participants

I interviewed 21 students currently studying at high school. Of these, 19 were currently enrolled in online high schools and two students were considering making the switch into an online high school. The biggest group of students came from Stanford Online High School (15) followed by Florida Virtual (2). Students varied in age from 14 to 17 years old. Recruitment was conducted through "snowballing" in which initial students referred additional students who could participate in the trial. Participants were asked to refer additional participants before any interviews were conducted to minimize participant self-selection in who they invited to participate.

While no global ranking exists for online high schools, Stanford Online High School would logically be the world leader in quality outcomes as online high schooling is most developed in the USA. It is also notable because it is currently ranked in the top 10 of all schools in America (including brick-and-mortar high schools) on ranking platform Niche (2020).

Procedures

I primarily interviewed students from Stanford Online High School consisting of 15 of the 21 interviews conducted because the school is the highest ranked online high school in America.

The 21 interviews were conducted through video-conferencing software without video cameras on and the interviewee was asked the survey questions outlined below under "Summary of Interview Questions". Interviews were recorded and transcribed.

I then conducted five deep sixty to ninety minute interviews with full transcriptions of the conversations with five of the respondents from the initial survey. Respondents were chosen by availability for a more extensive interview on a first

come first serve basis. Interviews were conducted online via a video call and recorded with the consent of the interviewee for further analysis using NVivo, a qualitative analysis software.

Analysis Methods

Thematic analysis was used to analyze the findings from the qualitative interviews. Thematic analysis is a technique used for identification and analysis of patterns in a dataset (Braun et al, 2006). It is used to isolate specific themes that are important to the phenomena in question (Daly et al, 1997). A primary strength of thematic analysis is that its flexibility allows for consideration of different types of data, in this instance, interview transcripts, that cannot easily be assessed through quantitative procedures. By clumping in themes, important factors can be isolated in an attempt to understand the underlying mechanisms that may drive a certain relationship (King et al, 2004).

In order to conduct the thematic analysis more rigorously, call transcripts from the deeper, exploratory interviews were uploaded into NVivo analysis software. Nodes were selected based on discussion points that emerged from the deep-dive interviews. Using these inputs, I produced a hierarchy tree-map which helps to visualize the relative frequency of different node occurrences. Each section of the tree-map is sized based on the relative frequency of occurrence both at the node and subnode level.

4.3 Weaknesses in Methodology

As with any surveying sample, there are a variety of biases to be aware of in this sample. Firstly, I am speaking to students who currently attend Stanford Online High School. These students are likely to have had an overly positive experience

compared to a universe of all students that have interacted with the school. This is a function of survivorship bias as students who did not have a positive experience may have opted out of the school already and left to brick-and-mortar high schools.

Secondly, Stanford Online High School is a global leader in online high-schooling and as a result, students are likely to be substantially more satisfied and also be of higher academic motivation than a traditional online high school student. This is clear given all students interviewed were high-achieving and ambitious, often juggling ambitious academic and extra-curricular schedules.

Thirdly, the students may feel uncomfortable sharing negative experiences about their current school. Even though students are kept confidential and the information is not shared with the school, they may be scared of information leakage that could impact their odds of being awarded leadership roles or other benefits inside the Stanford Online High School community.

Fourthly, the sample size of 21 is reasonable for a qualitative investigation but is not sufficiently large to be considered statistically significant.

Fifthly, students are generally clumped in the final years of high school so the survey is skewed towards more mature students who are potentially relatively more equipped for success in an online high school with more developed social skills than younger students. I did offset this partially by a variety of interviews with younger students.

Sixthly, Stanford Online High School is a synchronous delivery model with live-classes in which students attend lessons with a peer and teacher and attendance is effectively mandatory. This is significantly different to asynchronous delivery models where students consume recorded video lessons on their own time and complete assignments without much peer to peer engagement which occurs in some forms of online high schooling. While these caveats are important to understand, the qualitative insights were valuable in adding contextual understanding to findings in

the systematic review that was conducted and the insights from the cross-sectional analysis conducted on online one-on-one interactions.

4.4 Role of the Researcher

My role as a researcher in this context is to try and tease out the thoughts and experiences of the high school students without biasing their insights or feedback. In order to do this, I used structured interviews, followed by unstructured interviews. This process limits the ability of my own biases from previous involvement in the education industry to contaminate the initial thematic analysis when collecting the sizable proportion of the interview data. During the unstructured interviews, my role as the researcher evolves in attempting to tease out findings from the participants while also seeking to safeguard participants from any challenging or uncomfortable memories or experiences. There is more scope for participant discomfort in the unstructured interviews as some of the underlying drivers of the decision to join an online high school are probed and explored. I have to exercise discretion in examining these drivers while also respecting the needs of the participants.

If the interview participants were made aware of my role leading an education organization, it may lead to biased responses because of the perceived power differential. As a result, in order to minimize potential bias, I introduced myself as a DPhil researcher looking at online high schools.

4.5 Summary of Interview Questions

I asked the following interview questions across my student interviews in order to understand consistent themes between students. These questions were designed after conducting training with Ann Porteus at Stanford University in the class "Designing Education Surveys" (EDUC 399A) to minimize survey bias in question

design. Relevant research which informed the design of these research questions included "A Survey to Assess Student Perspective of Engagement in an Active-Learning Class" which leveraged a mixed-methods iterative design approach to design a survey after cognitive testing and exploratory factor analysis (Wiggens et al, 2017). The questions were also designed with the context from "Measuring What Matters" which emphasized the need for contemporary theoretical conception of feedback as opposed to explicit survey questions addressing student satisfaction and perception which can create unintended bias (Winstone et al, 2019). As a result, my survey questions do not directly ask about student satisfaction but err towards being objective and dispassionate with a focus on explicit academic outcomes and students' decisions to enhance the objectivity of response.

1. In what type of school are you currently enrolled (public, private, homeschool, blended)? What led you to make this choice? 2.[If online high school] Why did you initially decide to use an online high school?
2. [If online high school] How does your desired enrollment (single course, part-time or full-time) at an online school fit into your broader academic plan and goals?
3. [If online high school] What impact did enrolling in an online high school have on your academic plan and goals?
4. How much time do you spend on academic subjects outside of the classroom?
5. How are your time management skills?
6. Are you able to work independently?
7. How would you describe your personality?
8. What factors determine whether you perform academically?
9. How is your connection with your teachers and/or tutors?
10. What level of involvement do your parents have in your education?
11. What factors (if any) inhibit your ability to learn effectively?

12. What are the biggest differences to you between the online learning environment and the traditional learning environment?
13. Do you think online school is better than, equal to or inferior than a traditional brick-and-mortar school? Why?

In designing these questions, I focused on neutrality and ambiguity in the perspective of our research to avoid biasing the responses of students. A potential risk factor to my research would be that I primarily interviewed students in relatively high-performing online high schools (Stanford Online High School and Florida Virtual). In saying this, my qualitative research is designed to provide contextual insight into findings I have uncovered from our systematic reviews into student outcomes and student satisfaction in online schools as opposed to trying to provide causal insights.

4.6 Results of key themes from qualitative interviews

(1) Self-motivation

Being highly self-motivated was a requirement outlined by more than half of the interviewed candidates. Most of these participants had quite explicit goals. One interviewee stated "I'm just wanting to, I guess, get into a good college. Yeah, or being able to choose what I want to do". The need for self-motivation permeates multiple aspects of success in online schooling. Firstly, strong self-motivation is required to proactively build relationships for teachers. In a physical environment, small questions can easily be answered before or after class with a brief encounter but such an opportunity doesn't readily exist in online schools. Instead, teachers often have structured office hours, which requires planning and motivation to attend and

ask questions. Another interviewee stated "I know... people that are self-motivated, including myself, and it is definitely so much easier to enjoy the online experience when you have that self-motivation". Additionally, online schools generally tend to have less proactive student management and put a higher burden of accountability on the student. As a result, self-motivation is necessary to avoid missing assignment deadlines, meeting course requirements and achieving strong academic results. A student said "I'm generally motivated... I like to finish tests as well as I can... I still feel bad if I turn into something that isn't up to a certain standard".

In a traditional school, social relationships between teachers and students also drive higher accountability as the level of discomfort a student may feel disappointing a given teacher is likely to be higher given the stronger bond. This informal accountability mechanism is reduced in traditional online schools resulting in more room for students to miss deadlines through procrastination.

The social experience of online schools also requires self-motivation to engage comprehensively. In a traditional school, social interactions can happen accidentally or with minimal friction or pre-arrangement. In an online schooling environment, proactive meetings have to be arranged via video call to facilitate social interactions. This activation energy is higher and subsequently is likely to be a barrier for introverted and less highly motivated individuals.

(2) Desire For Academic Acceleration Through Extension Coursework

A desire for academic acceleration was a requirement outlined by more than half of the interviewed candidates. Most candidates cited reduced availability of coursework at traditional brick and mortar schools, limited access to specific curriculum (such as advanced placement) or limitations on competency based learning pathways. One student remarked "Next year I am taking honors chemistry, honors intermediate

algebra, history and philosophy of science, AP Spanish, critical theory and Russian literature". Very few of these courses were available at traditional offline schools near the student. Another student remarked they were taking "AP Microeconomics and Advanced Microeconomics as well as Advanced Biology and AP Calculus BC". Another student said "with public school, they try to keep people at the same level. So they don't let people get ahead in their grade and they don't let you skip classes, but in online high school, you will kind of determine what grade you are based on what courses you are able to handle".

Traditional brick and mortar schools with salaried teachers and a higher share of fixed costs tend to have stricter requirements on offering a broad range of classes. Classes will subsequently require a higher number of students to be economically sustainable given elevated teaching costs as well as physical real estate. Online schools, with a lower cost basis (~40% below traditional schools (William Blair Equity Research, 2019) subsequently can sustainably offer a higher number of classes.

Beyond offering more classes to choose from, students can typically opt to take a higher than normal course load in terms of quantity. Often traditional schools given rigidity on their local supply given requirements to be able to physically get to a school in a given day and a higher cost basis tend to restrict students to a standardized number of classes, irrespective of academic ability or offer minimal flexibility. A number of students remarked they were taking substantially harder loads than typical students at brick and mortar schools as well as accessing university content.

Generally schools tend to have limitations on what curriculum they can offer. Online schools generally face similar requirements to be able to offer a given curriculum but larger scale (in terms of student volume) can make it economically feasible to offer more curriculum offerings. A number of interviewed candidates remarked that it was difficult to find the advanced placement coursework available

in their online school in local traditional schools.

Finally, schools in an attempt to standardize the student experience and reduce the risk of student failure generally require students to progress through each academic year based on their actual age. One consistent theme from interviewed candidates was their attraction to the "competency" based progression framework at online high schools where students are able to progress through academic levels based on ability rather than age. This unlocks the opportunity to participate in university level coursework while in high school or space out completion of challenging advanced coursework (like advanced placement or A Levels) over several years. Competency-based learning that decouples age from academic progression is enticing to self-motivated students and unattractive to unmotivated students as they lose the pseudo-automatic progression to the next year.

(3) Less Stressful Social and Competitive Environment

Four of the interviewed candidates cited the mental health benefits of online schooling in reducing social pressure in being able to ask questions, navigating broader interactions with peers, helping to grapple with Attention Deficit Disorder (ADD) and reducing perceived peer competition. One student remarked "the place I came from [before online high school] had a super toxic academic environment. So instead of like, other people helping each other out, it was always if you get a higher grade than me I am going to put you down. It was already starting to be every person for themselves in middle school. So I don't know how I would have dealt with that in high school".

Typically in a school environment in which students know one another deeply, the social cost to asking a question that is perceived to be ignorant or ill-informed is high and in equilibrium, many students in a class may not ask a question despite not understanding the content being presented. In an online environment in which

peers have a weaker relationship and are not physically co-located to provide social recourse to questions straight after class, the fear associated with asking a question declines and more students are able to participate.

In most school environments, a social hierarchy tends to naturally become established with different cliques often unified by shared interests, background or reputation. Given the lack of out of class interactions in an online school, the capacity for this social hierarchy to develop is greatly reduced which reduces social anxiety. Additionally, the ratio of social interactions in an online school tends to feature more one-to-one calls as opposed to group calls. One-to-one calls tend to reduce the prevalence of peer pressure and harassment as compared to group interactions so this shift in the communication norms further reduces social pressure.

Finally, in a traditional school environment the diffusion of knowledge around student performance in academic assessments and informal discussion outside of class is substantially higher than an online school. As a result, online school generally tends to result in less competitive pressure with peers. Additionally, the online school's focus on competency based learning reduces the ability for peers to directly compare courses with one another as they have varied ranges of classes depending on their performance with a given subject. The social expectation to share grades is reduced in an online school because examination and test results tend to be distributed electronically and therefore, more privately, than a traditional brick and mortar school, which may hand out assessments in public settings. The reduced competition may have a first order effect of reducing student quality but may result in second order effects such as higher intrinsic motivation or higher perceived goals.

(4) Flexibility in Schedule Facilitating Other Activities

A recurrent theme in interviews was the strong perceived benefit of the flexibility online school offered. Students from varied backgrounds including gymnastics,

fencing, tennis and dancing cited the infeasible constraints of brick and mortar school schedules with their training regimes. Online schools offered options to take courses around the gaps in their training regime to stay on top of their educational journey without compromising on their other pursuits. Beyond timing flexibility of courses, which are often offered at multiple different times throughout a week, the use of online delivery removes the need for geographical limitations on where the student is based. For a high performing tennis player, for example, who seeks to compete in an international circuit, the ability to log into classes from around the world enables a substantially higher attendance rate than a physical school. One student interviewed remarked that "I was able to balance my 20 hours/week of dance commitments alongside the flexibility of online school".

4.7 Perceived Weaknesses of Online Schooling

(1) Higher Barrier for Communication with Teachers

The majority of interviewed candidates commented that online schooling makes it more difficult to build strong relationships with teachers. In traditional brick and mortar schooling, students have substantial informal interactions with teachers outside of class time, which builds rapport. Additionally, students can easily ask questions outside of class, which helps teachers build a relationship with a student and potentially feel more invested in the relationship. In an online school, students have to intentionally choose to attend office hours to have teacher interactions. Additionally, there tends to be a reduced presence of assemblies or extra-curriculars in which teachers and students often build relationships. The compounding effect of not asking lots of relatively simple questions that may get asked in a brick and mortar environment is a substantial reduction in total time spent speaking with teachers outside of class which according to interviewed candidates, results

in a weaker core relationship that needs to be fostered by more proactive steps. Additionally, teachers play less of a role in chasing students as a derivative of the lack of interpersonal interaction. One student summarized this as "in a public [physical] school, you have your teachers there and they are constantly reminding you of stuff and get on your ass if you don't turn in an assignment. If you miss stuff in an online school you have to take responsibility yourself. You don't feel that same kind of accountability to your teachers".

(2) Less Spontaneous Opportunities for Social Interaction

Similar to (1) with limitations in building relationships with teachers, students typically interact in a traditional school through extra-curriculars, lunch and afterschool activities and informal meetings between classes. In an online school, the focus on contact with students primarily being group classes and then some limited online extra-curricular activities tends to meaningfully reduce the average number of social interactions of a typical student. Highly motivated students are able to proactively build relationships with other students but the limitation to organic interactions is likely to affect more introverted students. One of the students remarked "I do think [the lack of social interaction] is a problem. Being an online high school, it is hard to break into certain social groups and it can be awkward sending someone a message as opposed to talking to them face to face".

(3) Limited Access to School Extracurriculars

Naturally, many extra-curriculars take place through co-location, particularly sports and so online school students generally face limited options or have to pursue out of school activities. School extracurriculars are a main channel for deepening social interactions, failing in front of peers in high stakes environments (like sports) and serves to provide students a broader view of their peers than their academic

performance alone. While online high school students can engage in extra-curriculars in the physical world with other students outside of their online school network, there may be some benefits lost from interacting in extra-curriculars with the same students that one does academic classes with. Online schools can partially mitigate this by introducing online extra-curriculars such as online debating clubs in which students can engage in live interpersonal interactions on global issues that are not directly related to classroom content. This has limitations when it comes to musical instruments or robotics and engineering. Additionally online schools can form partnerships with community organizations like sports clubs, physical schools that have orchestra, theatre and other offerings and other groups to provide an extra-curricular suite for students.

One student said "I was asked what I was going to do about social interaction? Firstly, I was still able to do a lot of extracurriculars. In those extracurricular activities, I think I have a very healthy social life. I see people everyday. I think academics is what I want to be best at. It is okay to put your social life second and your academics first. I can still do a lot of extracurriculars online and the ones that I can't do online, I can just participate in with local groups generally."

4.8 NVivo Analysis

As described in the methods section, I conducted five deep sixty to ninety minute interviews with full transcriptions of the conversations with five of the respondents from the initial survey. Respondents were chosen by availability for a more extensive interview on a first come first serve basis. Interviews were conducted online via a video call and recorded with the consent of the interviewee.

Call transcripts were uploaded into the NVivo analysis software. I constructed the following node map:

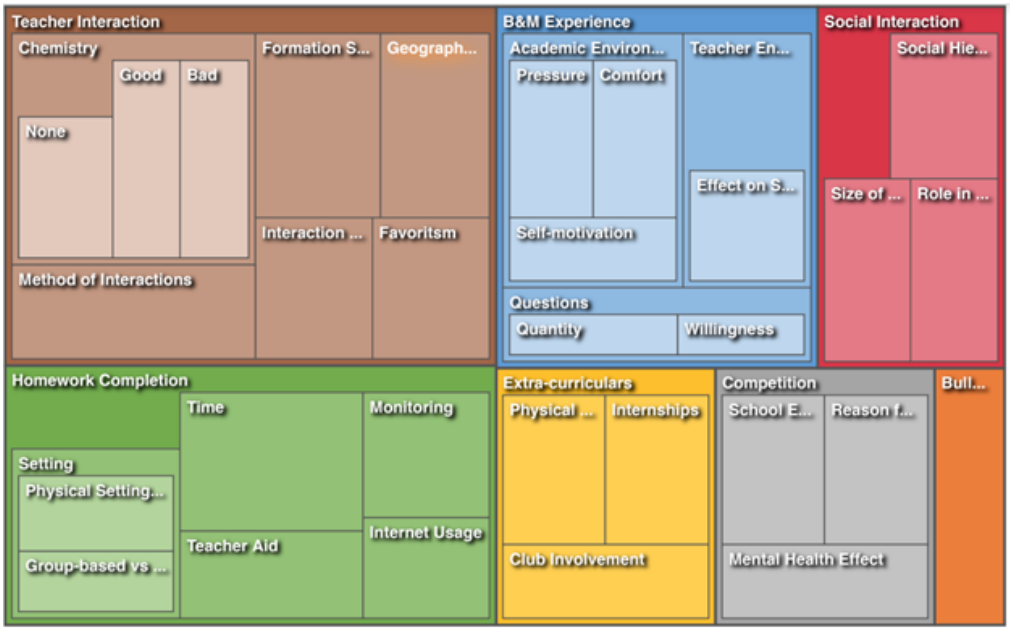
Nodes

Name	Description	Files	References
B&M Experience	(Node Label)	0	0
Academic Environment	(Node Label)	0	0
Comfort	Student's self-perceived comfort level in academic environment	5	5
Pressure	Pressure student feels in academic environment	5	5
Self-motivation	Amount of self-motivation student perceives in academic environment	4	4
Questions	(Node Label)	0	0
Quantity	The quantity of questions students raise	4	4
Willingness	Willingness to ask questions	3	3
Teacher Engagement	The energy and engagement level of teachers	5	5
Effect on Student	How a teacher's engagement in the class affects the student	5	5
Bullying		5	5
Competition	(Node Label)	0	0
Mental Health Effect	The mental health effect of academic environment's competitiveness	5	5
Reason for Competitiveness		5	5
School Environment	Nature of school environment as described by student	5	5
Extra-curriculars	(Node Label)	0	0
Club Involvement		5	5
Internships		5	5
Physical Activity		5	5
Homework Completion		3	3
Internet Usage	Amount of internet usage by students for homework	4	4
Monitoring	Parent Monitoring if any	5	5
Setting	(Node Label)	0	0
Group-based vs individual setting	Group-based versus individual setting for completing homework	4	4
Physical Setting (at home vs a cafe)		5	5
Teacher Aid	Teacher aid for homework	5	5
Time	(Duration of time for completion)	5	8
Social Interaction		2	3
Role in Friend Group		5	5
Size of Friend Group		5	5
Social Hierarchy	The level of social hierarchy as perceived by the student and their friend group	5	5
Teacher Interaction	(Node Label)	0	0
Chemistry	Chemistry between teacher and student	2	2
Bad	How often students had bad chemistry with teacher	5	5
Good	How often students had good chemistry with teacher	5	5
None	How often students had no chemistry with teacher	5	5
Favoritism		5	5
Formation Setting	Settings that students formulated relationships with teachers	4	7
Geographical Proximity		5	6
Interaction Difficulty	Difficulty of interacting with teachers as perceived by students	4	5
Method of Interactions	How students bonded with their teachers	5	7

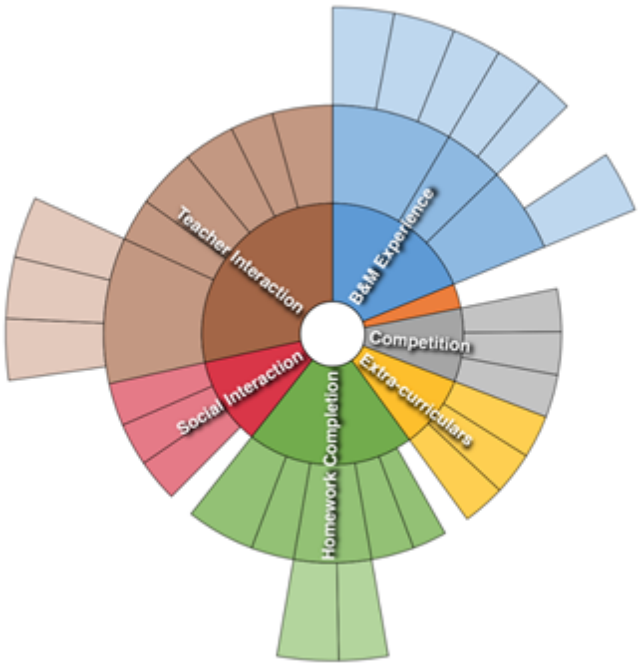
4. Qualitative Analysis

Nodes were selected based on discussion points that emerged from the deep-dive interviews. The column "File" tracks how many of the respondents discussed the node in question ranging from 0 to 5 responses. This would be the equivalent of five if(x>0,1,0) commands in excel that are binary indicators that are switched on if the node is engaged with. The column "References" tracks the total number of references to the node in the interviews. By design, "References" has to be equal to or greater than "File" because an individual who engages with a node topic will be captured only once in the "File" column but can appear several times in the "References" column. The description outlines a more explicit definition of each "node" name. This helps to provide detail as to in what context each node was discussed.

Using these inputs, I produce the following hierarchy tree-map which helps to visualize the relative frequency of different node occurrences. Each section of the tree-map is sized based on the relative frequency of occurrence both at the node and subnode level. For example, pressure and comfort within the academic environment both have 5 occurrences so are equally weighted in this visualization. Quantity appears 4 times under questions and willingness appears 3 times which is why the quantity visualization is ~33% bigger than the willingness node. The hierarchy charts essentially show if some nodes have more coding references than others to identify prominent themes in the analysis.



This visualization helps to articulate the depth within each node:



The NVivo analysis reinforces many of the same themes as the initial interviewing efforts. I now discuss some of the additional insights that were highlighted.

Firstly, monitoring in the form of recording of classes has various behavioral effects on students and teachers. Students cited that they observed generally stronger

behavior in their online learning environments than in traditional brick-and-mortar classes. One student noted that it was harder to hide because the teacher can see every student, can see if one is not focused on the screen or task at hand and any inappropriate behavior can be replayed and scrutinized. Another student noted that teachers are often significantly more focused during online schooling sessions and are punctual to class than experiences in traditional brick-and-mortar schools. This is logical if the ability to use recording provides the headmaster and other school administrators clear ability to detect teachers who are consistently late to class, wasting class time or acting inappropriately.

Secondly, the concept of hierarchy within the online school environment was discussed. Interestingly, at Stanford Online High School students said that those individuals who lived closest to the physical campus and could subsequently attend in-person meet ups and also socialize in the physical world more often tended to have relatively stronger positioning in the informal social hierarchy within the school. This could be explained by the students who have built physical relationships having a higher propensity to talk to one another in the online environment and subsequently a higher share of overall class time and social interactions flow through them. It may be interesting to consider fully online meet-ups to offset the ability for this effect to grow if it becomes problematic and imbalanced. Additionally, hierarchy from the perspective of age is challenged with the online high school environment because classes are streamed by competence so students can sit in a class with younger and older peers. Relative hierarchical power tends to accumulate around individuals who are high achieving for their age as opposed to the students who are oldest in a relative scale. In a traditional school where older students consistently do more advanced work, age is a more fitting proxy for ability but in a streamed, online high school like Stanford this is not the case.

Thirdly, bullying was analyzed in greater detail. An interesting observation

was pointed out that in an online school, all types of social interactions may tend to occur less frequently so it is natural that bullying also decreases. There may not be anything intrinsically different about the students in the online high school that makes them less inclined to be bullies, but with less opportunity for social interactions that constitute bullying, the incidence declines. It was also shared that most bullying would take place on platforms outside of the school communication infrastructure, presumably because it could be less easily tracked. Some of the literature covered in the systematic review suggested the higher proportion of time spent online in an online high school may result in higher rates of cyberbullying but this did not seem to reconcile with the commentary of Stanford Online High School students.

Fourthly, favoritism dynamics was discussed in the long-form interviews. Two students independently argued that favoritism is more easily detectable in the online high school environment where participation is a mandatory part of grading because students feel pressure to contribute to class discussion and people are subsequently more attuned to an individual being repeatedly called on. On the flipside, in a traditional classroom the only way of answering a question is usually with one's voice but in the online high school, instant messaging on the chat function helps many classmates respond simultaneously either in a group or privately to the teacher. This helps to offset dominance of one or two students crowding out engaging opportunities for other students.

Fifthly, students created opportunities for serendipitous social interactions. It was common in the culture of the Stanford Online High school that students would have large group Skype calls that students would join when they had available time and wanted to socialize, similar to hanging out in a hallway or common space area where you may bump into other students in a physical brick and mortar school. Students also noted that there was a higher degree of proactivity in scheduling calls

between students to meet one another than the more informal peer matchmaking processes that occur in a physical brick-and-mortar school. It was quite normalized to "connect" and have exchanges where both students introduced themselves without any particular agenda or reason for the interaction. This was an important source of social bonding that occurs outside of traditional classroom hours.

Finally, extra-curricular activities need to be more transparently updated in an online high school. In a physical school, signs around the school, groups of students walking in the same direction at lunch time and spontaneous interactions provide informal information about what activities are happening around school. In an online high school, it is more easy to miss these types of cues and as a result, it is potentially harder for students to be aware of what extracurriculars are active and occurring than in a physical brick-and-mortar high school. As a result, it was noted to be important that schools maintain an active list of available extracurriculars, ideally with participant numbers and clear information on how to join. Students were also more willing to be sent a higher volume of information through email to avoid missing out on potential opportunities than in brick-and-mortar schools, according to an individual who had participated in both full-time online and offline schooling environments.

4.9 Discussion

In this section, I will discuss some of the key findings from my qualitative analysis and how these insights fit into the broader existing academic literature on online high schooling.

Firstly, the theme of online high school students thriving if they have strong self-motivation levels is broadly supported by the literature. High Enrollment Course Success Factors in Virtual Schools: Factors Influencing Student Academic Achievement (Liu et al, 2011) found a high correlation between time spent engaging

with the online content inside the learning management system and academic outcomes. Students who have the discipline and determination to engage with the material outside of class time are naturally more likely to be successful. This is rather intuitive as in an environment like online high schooling when the student generally has more flexibility and some of the "safety bars" of a traditional high school are removed, students who skew towards being more self-motivated are more likely to thrive.

Secondly, the reduced opportunity for social interaction which was discussed by a significant number of interviewees is well supported in the literature. An Investigation into Reported Differences Between Online Foreign Language Instruction and Other Subject Areas in a Virtual School (Oliver at al, 2012) found that subjects like languages which generally require more peer interaction are harder to teach in an online learning environment. Serendipitous interactions between students are likely to directly improve learning outcomes for language acquisition if they are practicing the target language. This is less likely to be true for something like mathematics where the mere act of talking doesn't necessarily improve a given student's performance.

Thirdly, my findings reconcile quite tightly with Perceived Advantages and Disadvantages of an Online Charter School (Shoaf et al, 2007). This paper involved running focus groups with 44 students, parents and teachers. The six most consistent themes were that virtual schooling provided more appropriate pace of instruction, provided greater access to special subjects that were difficult to find in traditional brick-and-mortar environments, offered greater access to individualized instruction, facilitated easier modification of lessons, enhanced flexibility in structuring the school day but resulted in limited social engagement.

My analysis found very similar themes with many of the students taking advantage of the greater access to special subjects that online schools offered

to embark on academic acceleration. A number of the students remarked about the flexibility of their schedule enabling participation in competitive sport and other activities which aligns with the findings of the paper that online schools generally provide enhanced flexibility. The theme of less scope for social engagement was also apparent across both papers. The consistency of my findings, despite the relatively overachieving segment of students interviewed, with the more general pool of students in Shoaf's analysis adds weight to the strength of these findings.

One paper that varied in part from the findings of my qualitative analysis was "An Online High School "Shepherding" Program: Teacher Roles and Experiences Mentoring Online Students" (Drysdale et al, 2014). This paper highlighted that the barriers for engagement with teachers was actually lower (rather than higher) because of the prevalence of alternative contact channels online such as google chat, text messages and voice recordings. The findings can be reconciled by acknowledging that while it may indeed be true that there are more ways to contact a teacher in an online environment which may potentially broaden the range of scenarios in which a student and teacher can build a relationship, one of the most common instances of engaging with teachers, namely bumping into them before or after class, is made harder in an online environment. It also may require relatively more discipline to set up a time with a teacher in an online context than in a physical environment when hanging around after class. It may be interesting for future research to examine the different channels a student in a brick-and-mortar school interacts with teachers and their relative prevalence and contrast this with an online high school environment. The analysis could consider both the breadth of channels and the volume of communication through each channel to attempt to tease out in which of the two environments more net interaction occurs between the student and the teacher.

Cyber Charter Schools: Evolutions, Issues and Opportunities in Funding and

4. Qualitative Analysis

Localized Oversight (Ellis et al, 2008) makes the argument that some of the flexibility of online schooling actually needs to be reigned in, in order to enhance student outcomes and reduce the level of self-motivation necessary in the child for the format to be successful. From my qualitative analysis, heavy restrictions on the flexibility of an online schooling environment like Stanford Online High School in which the student body are very self-motivated may be unnecessarily punitive given the students tend to benefit from the flexibility of the school through broadening their scope for participation in extracurricular activities and don't appear to be exploiting the flexibility to avoid their academic responsibilities as Ellis and his colleagues argue.

Students with Special Health Care Needs in K-12 Virtual Schools (Fernandez et al, 2016) builds on my findings that online high schools can be a safe-haven of sorts for students with reduced instances of bullying, social hierarchy and intensive academic and social stress. In this study, students with special health care channels didn't perform worse in an online environment than a physical environment. While it would be more compelling if the paper found that the students overperformed in online schools, it is interesting that they performed at a comparable level despite the potentially higher need for self-motivation in an online high schooling environment. My qualitative analysis found a common theme of there being less stressful social dynamics and a less stressful academic environment observed by the online high school students interviewed.

In conclusion, my qualitative research generally reconciles with the existing academic literature and in some areas, builds on the existing understanding. I was able to build on existing work by strengthening the understanding of the drivers for participation in an online high school environment such as seeking academic acceleration and looking to avoid the stressful social and academic environment in some traditional high schools. I was also able to tease out an interesting nuance about social hierarchy construction in online high schools in which it appears that

the students who are more able to meet up physically on a more frequent basis tend to build a stronger social network and as a result, be more "popular" within the online high school environment. This would suggest it is important that offline social networking opportunities are made broadly available. Additionally, I was able to dig more deeply into how some of the students compensate for the lack of traditional social interactions through things such as scheduled group video-calls or hang out call links in which people just in when they have free time to spend time with their classmates. While not analyzed in my work, emerging digital tools such as Slack and Discord appear to have amplified the capacity for social interaction in the digital environment which may reduce the perceived social interaction gap between physical and online schools relative to early literature in this space. The mechanisms developed in this qualitative analysis and my systematic review of the drivers of student outcomes and student satisfaction form a strong foundation for my public policy recommendations in chapter seven of this thesis.

5

Cross-Sectional Analysis

5.1 Background

In this chapter, I conduct a cross-sectional analysis to identify the relationship between the psychometric traits of students and teachers/tutors and the resulting student satisfaction and cancellation rates in an online learning environment. Psychometrics are defined as the "psychological technique of mental measurement". After seeing the gap in the literature on education matching around cognitive style and personality, this chapter plays an important role in exploring some initial potential relationships and seeing whether any correlational effects may exist. The cross-sectional analysis is also a relatively more time-efficient method of assessing the potential of psychometric matching before embarking on the full randomized control trial later in the thesis. The context from this chapter helped to design

the randomized control trial experiment.

The relationship between the teacher and the student is a complicated interpersonal dynamic that requires trust, accountability, patience and other dynamics. Teaching styles and cultural norms around student-teacher interaction vary significantly from region to region and even from school to school. It is possible that certain combinations of student and teacher pairings could achieve stronger academic outcomes and satisfaction than random matching of student and teachers based on factors like cultural background, learning needs, predisposition for formation of a strong relationship (based on psychometric traits) and other interactions. Psychometric profile targeting has been used in political contexts to more effectively tailor messaging to be received by prospective voters (Dobber et al, 2020). Psychometric profile targeting has also been used in friend recommendation algorithms on Facebook and other platforms (Bian et al, 2011). While the use of psychometric matching in education is limited, its effectiveness in other formats to build strong relationships and enhance the "stickiness" of a message and information would appear relevant to the classroom (Heath et al, 2007).

In chapter three, I conducted a systematic review looking at existing research into the potential for matching to improve student outcomes or student satisfaction. The primary types of matching studied in the literature involved ethnicity, gender and cognitive-style matching. One of the areas I identified as having potential for further analysis was cognitive-style matching. This was a category of matching that had limited academic research but some promising results (Packer et al, 1998). Although cognitive-style and personality are not the same thing, they do have a relationship: *"In essence, cognitive styles are characteristic self-consistencies in information processing that develop in congenial ways around underlying personality trends"* (Messick et al, 1984). It was challenging to obtain a dataset of meaningful size that captured cognitive style. I was able to source a dataset that featured

psychometric personality traits which offers a reasonable foray into various attempts to quantify specific characteristics of students that may lead to personalized learning interactions (such as in mentor selection).

This chapter utilizes a large-scale data-set from an education organization, Crimson Education, that captures personality characteristics of both students and tutors as well as the resulting student satisfaction scores following one-one-one tutoring lessons. This is important because it enables some empirical assessment of the potential for personality-trait matching to improve student satisfaction and considers a potentially useful matching criteria that has been scarcely addressed in the existing related work found in the systematic review. While no data-set could be sourced that specifically captured cognitive-style matching, my analysis using HEXACO personality trait matching offers a relevant contribution to matching literature that has not been sufficiently explored. Personality is likely to be an important factor given the expansive literature around the interpersonal dynamics between student and the teacher as well as teacher expectations of the student being important contributors to student outcomes (Kwok et al, 2007). A personality trait such as conscientiousness which measures how reliable and consistent the student is would appear to be relevant to considerations like if homework will actually be completed, how attentive the student will be in class and what types of teaching strategies might be more effective.

Student satisfaction could be affected by various factors including their perception of the learning experience with their teacher (including how they feel after the lessons and throughout the learning process) or the objective outcomes of the learning experience (did the student acquire some new content they now understand?). In order for the personality of either the student or tutor to affect student satisfaction, there are three types of interactions that could be relevant. The tutor's psychometric traits could affect the student's satisfaction levels. The

student's own psychometric traits could affect their satisfaction levels. Additionally, some interaction that could potentially be predicted based on the psychometric traits of the student and the tutor could occur that could affect satisfaction levels.

In general, if the relationship between psychometric traits is important between students and teachers, these trait differences should either be maximized or minimized in a simple statistical model (Hottung et al, 2019). Derived from this logic, I design a number of hypotheses based on in what direction I would expect the relationship between psychometric characteristics to impact student satisfaction positively.

In order to test these differing mechanisms, I consider the impact of tutor psychometric characteristics on student satisfaction score. Secondly, I consider the impact of student psychometric characteristics on student satisfaction score. Thirdly, I consider the impact of the relationship of the student and tutor psychometric characteristics on the final score. Fourthly, I conduct some additional analysis into the impact of psychometric characteristics on cancellation rates which are likely to be an endogenous factor in observed student satisfaction rates. Finally, There is some evidence to suggest cultural attitudes towards education varies, teacher qualifications, respect for the status of the teacher, competitiveness of the education system, spending on private education and the number of hours spent on learning all vary by region which may result in regional idiosyncratic differences (Kaarsen, 2014). I use a global data set to maximize generalizability.

5.2 Description of Relevant Psychometric Traits

People with very high scores on the Honesty-Humility scale avoid manipulating others for personal gain, feel less temptation to break rules, are less interested in lavish wealth and luxuries, and feel less entitlement to elevated social status. Conversely, persons with very low scores on this scale are more likely to flatter others to get what they want, are inclined to break rules for personal profit, are motivated by material gain, and feel more self-important (Ashton et al, 2009).

The sub-trait inquisitiveness assesses a tendency to seek new information and experiences. For example, low scorers may have little curiosity about the natural or social sciences, whereas high scorers may read widely and be interested in new experiences like travel (Ashton et al, 2009).

People with very high scores on the Extraversion scale feel positively about themselves, feel more confident when leading or speaking in front of people, enjoy social gatherings and interactions, and experience positive feelings of enthusiasm and energy. Conversely, people with very low scores on this scale consider themselves unpopular, feel awkward when they are the center of social attention, and are less likely to enjoy social activities (Ashton et al, 2009).

The more extraverted the student is, the more likely they are to have spontaneous social activities or other last minute obligations that are a higher priority than their tutoring session resulting in an elevated cancellation rate. They are also more likely to want to attend any events that are occurring given the heightened utility they derive from large gatherings over a one on one interaction such as tutoring.

People with very high scores on the emotionality scale experience fear of physical dangers, experience anxiety in response to life's stresses, feel a need for emotional support from others, and feel empathy and sentimental attachments with others. Conversely, people with very low scores on this scale are less deterred by the prospect of physical harm, feel little worry even in stressful situations, have little

need to share their concerns with others, and are less likely to be emotionally dependent on others (Ashton et al, 2009).

Students with low emotionality scores may feel less need to have the reassurance of a tutoring session and rely less on sharing frustrations with other individuals and in my analysis were marginally more likely to cancel a session. Additionally, if a student has a higher emotionality the tutor may fear adverse consequences of cancelling a session more and is thus more likely to stay committed to delivering all the intended sessions.

A tutor with a lower score for agreeableness is more likely to get frustrated with their student perhaps after a frustrating session or a lack of homework completion and subsequently cancel more sessions. A tutor who has a lower score emotionally is likely to be more emotionally detached to their student and view their role as more of a job as opposed to a sustained social obligation and subsequently cancel at a higher rate. Additionally, a student is less likely to feel a strong sense of attachment to a low emotionality tutor and is thus more likely to cancel on them without feeling remorse.

Tutors with a lower score for honesty and humility are more inclined to break rules and have a larger sense of self-importance so will put less weight on their responsibility to their students and rules around cancellation. A tutor who does not tend to follow the rules may seem less intimidating for a student to cancel a session with, however, this is not entirely coherent as the tutor may actually be more strict on student cancellations stemming from a desire to earn money from completed sessions when they want the session. Tutors with a higher score for conscientiousness would seem less likely to cancel but I actually observed a positive correlation. This may be because tutors that tend to follow the rules more are likely to report a higher proportion of cancellations properly through the Crimson application that manages booking. Additionally, highly conscientious tutors are likely to drive more accountability in the student to book and subsequently cancel

sessions through the Crimson application.

5.3 Hypotheses

I developed six core hypotheses that I sought to test through the cross-sectional analysis.

H1: Matching based on minimizing the absolute difference between trait conscientiousness of the student and the teacher will have a statistically significant positive effect on student satisfaction.

Thomas Duffy found that trust and learning intentions are important factors which impact student perception of the learning environment and potentially learner performance (Duffy et al, 2004). A student is likely to be more satisfied when they don't feel like they are letting down their teacher or vica versa so the relationship is not being stressed through missed expectations. Trust is more likely to be impaired when a difference in consciousness exists between student and teacher such that one party lets down the other because of a lack of consistency or reliability (in submitting homework, returning feedback on homework, attending class or other such components of the learning experience).

H2: Matching based on maximizing the absolute difference between trait extraversion of the student and the teacher will have a statistically significant positive effect on student satisfaction.

Existing research has not found any variance in preferred teaching methods in the classroom or differences in preferences for participation activities between introverted and extroverted students (Parkman et al, 2017). In "Extraversion and Introversion Personality Type and Preferred Teaching and Classroom Participation" (Nina et al, 2017), significant differences in students' preferences for "engagement in discussion with other students" were found when evaluating students with

different levels of extraversion.

Based on this work, I hypothesise a similar effect would drive a student's willingness to participate in active discussion with their teacher. If extroverted teachers are more willing to talk for a higher share of the classroom lesson, for example, it would reduce the need for an introverted student to proactively verbally participate. A student who is introverted is likely to be able to fall into their natural preference styles for communication if they are interacting with an extroverted teacher who can drive the conversation. An extroverted student is unlikely to mind whether the teacher is introverted or extroverted. Additionally, even if the teacher is introverted, given they have the authority role, they are still likely to teach effectively relative to an extroverted teacher as they are comfortable with the material and mentally equipped for their role (as compared to spontaneous, unstructured social interaction at an informal mixer, for example). Looking at the directions of these various hypothesized variable relationships, it would seem that matching based on the absolute difference in trait extraversion (or in simple terms, attempting to match introverts with extroverts and vica versa) is likely to improve student satisfaction.

H3: A positive correlation will exist between measures of student satisfaction and student outcomes.

Structural equation modeling conducted in the 2021 paper "Teacher Feedback Practices, Student Feedback Motivation and Feedback Behavior" found that there existed a positive relationship between student feedback and student outcomes in which student feedback significantly predicted course exam results (Gan et al, 2021). If these findings are generalizable to my data, it would imply a positive correlation will exist between measures of student satisfaction and measures of student outcomes.

H4: Matching based on minimizing the absolute difference between trait conscientiousness of the student and the teacher will have a small statistically significant positive effect on student outcomes.

5. Cross-Sectional Analysis

Based on H1 and H3, if I believe that student outcomes and student satisfaction are positively correlated and that matching based on minimizing the absolute difference between trait conscientiousness will have a positive effective on student satisfaction, applying transitivity logic will imply a positive effect on student outcomes when minimizing absolute differences between trait conscientiousness of the student and the teacher. Transitivity is a term from microeconomics which is an assumption in preference stability such that if a person prefers A over B and they prefer B over C, they should prefer A over C (Dana et al, 2011).

H5: Matching based on maximizing the absolute difference between trait extraversion of the student and the teacher will have a statistically significant positive effect on student outcomes.

As an extension of H2 and H3, I hypothesize H5 based on a similar transitivity logic (described in H4).

H6: Matching based on gender will have no statistically significant effect on student satisfaction or student outcomes in aggregate, although effects may vary for students of different cultural backgrounds.

While matching based on gender during early years of education may yield positive effects of role modelling (Cho et al, 2012), by the time the student is sixteen or older and are succeeding academically (as is generally the case in the sample), it would appear likely they have been able to successfully learn under both male and female instructors and some of the benefits of comfort from interacting from a teacher of the same gender would be abated. As a result, it would appear unlikely they would receive any particular boost from gender-matching. While one could argue this logic would also apply to personality, it may be that people of a specific gender have an internal mental model or assumptions for the personalities of individuals from the same gender and different genders that they initially hold which then breaks down and converges on the real personality characteristics of

the individual they are interacting with after spending some time together. This will be tested in Chapter Six (randomized control trial).

5.4 Measures

A leading personal structure is the Big-Five personality traits developed in psychological trait theory from the 1980s. The psychometric test being used is the HEXACO Personality Inventory developed by Kibeom Lee and Michael Ashton. This is an evolution of the Big-5 Personality Inventory that includes six dimensions: Honesty-Humility, Emotionality, Extraversion, Agreeableness, Conscientiousness and Openness to Experience and is designed for contexts in which there is limited administration time for the survey (as is the case in this type of experiment in which students may have limited attention span for extensive surveys). It adds Honesty-Humility, which is designed to make the assessment more relevant across varying cultural and language backgrounds (Ashton et al, 2018). A person's HEXACO score is produced after they answer 100 multi-choice questions which generally takes about twenty minutes, depending on the speed of the participant. The HEXACO assessment has become increasingly prevalent and in a large-scale meta-analytical assessment of the construct, 426 individual meta-analyses, 436 independent samples and 3893 effect size estimates were used to validate the efficacy of using HEXACO to map personality. The analysis demonstrated the internal validity of the assessment and a broad range of empirical tests which helped to authenticate its use for broad use of the HEXACO construct (Zettler at al, 2020).

5.5 Methods (Participants)

Context on the Sample

In total 69,779 sessions were conducted between student and tutor pairs. All students were enrolled in Crimson Education with a mean age of 16.2, a lower quartile of 15.8 and an upper quartile of 16.2.

The students primarily attend private schools or international schools and have the intention of leaving their country to pursue tertiary studies in the US or UK upon completion of high school. They generally are highly academic in the top 10% of their school's academic performance. The students (and their parents) have opted into purchasing tuition services from the platform that are 1-1 in nature and delivered online with the primary goals of increasing academic performance. The students are primarily European, Chinese, Korean and Indian with European being the most common demographic followed by Chinese. The students include both male and female representation in approximately equal proportions.

The teachers are primarily traditional college-aged students (aged 18 – 22) from university programs including Harvard, University of Auckland, Yale, Princeton and other comparable institutions. The teachers are generally not the same nationality as their students. They are contractors who typically engage with on average 2 students through the platform. The teachers are not accredited but have gone through a screening program and a training program before beginning their work with students and have applicable child-safety checks in place. They are compensated for their time based on the number of hours they interact with their students for. They generally stay with the same students throughout the course of the student's program except in instances of sickness or other challenging life circumstances. The teachers are primarily European, Chinese, Korean and Indian with European being the most common demographic followed by Chinese. The students include both

male and female representation in approximately equal proportions.

I excluded from the data ratings related to an old rating system from a previous version of the Crimson application used inside the company that had a different scale offering differing emojis to track a student's satisfaction instead of a clear 1 - 5 rating because variance in the scales and characteristics of the old rating system used would have made comparability of data questionable.

All of the variables being measured are routinely administered for all Crimson users by the company.

5.6 Methods (Statistical Techniques)

I ran 13 stepwise regressions to attempt to understand the relationship between psychometric characteristics, session score and cancellation rates. The main regressions I assess are:

1. The impact of student HEXACO factors (independent variable) on average rating (dependent variable).
2. The impact of student HEXACO factors (independent variable) on average rating (dependent variable) weighted by the square root of the number of ratings on each student-tutor pair.
3. The impact of tutor HEXACO factors (independent variable) on average rating (dependent variable).
4. The impact of tutor HEXACO factors (independent variable) on average rating (dependent variable) weighted by the square root of the number of ratings on each student-tutor pair.
5. The impact of the difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable).
6. The impact of the difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable) weighted by the square root of the number of ratings on each student-tutor pair.

7. The impact of the absolute difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable).
8. The impact of the absolute difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable) weighted by the square root of the number of ratings on each student-tutor pair.
9. The impact of cancellation rate (independent variable) on average session rating (dependent variable).

10. The impact of student HEXACO (independent variable) on general cancellation rates (dependent variable)*.
11. The impact of student HEXACO (independent variable) on tutor cancellation rates (dependent variable)*.
12. The impact of tutor HEXACO (independent variable) on general cancellation rate (dependent variable)*.
13. The impact of tutor HEXACO (independent variable) on tutor cancellation rate (dependent variable)*.

*refers to logistic regression. All other regressions are standard linear regressions.

The stepwise regression technique is criticized because of the fundamental challenge that some real explanatory variables that have causal impacts on the dependent variable may not be statistically significant, while other variables may appear to be statistically significant.

I chose this methodology because it is challenging to develop clear intuition using the HEXACO model as to precisely which personality traits should be included or excluded and what the directional impact of their effects are likely to be. This complexity is further compounded if one starts to consider models that use the relative or absolute difference between the personality dimensions of two participants. In order to compensate for the problem of picking up spurious correlations, I use the stepwise regression technique on a historical dataset (named "Pre-RCT Cross-Sectional Data") from Crimson Education with full knowledge of its deficiencies to build a robust model to test. I then take this econometric model and test it on a new sample obtained from a different time period at Crimson Education (named "RCT Data") during the randomized control trial (analyzed in chapter six).

As a result, rather than simply reporting statistically significant findings from the first sample (which may be spurious), they are then tested on a fresh sample to assess whether analyses replicate in an independent dataset. For something

fairly obscure such as personality where it is hard to develop intuitive models for why certain sub-characteristics (from a selection of 32) between two individuals may have a specific effect, this approach provides a higher likelihood of detecting subtle relationships between variables (such as the impact of matching people based on absolute difference in personalities).

The only exception to this is for cancellation rate data in which no data existed in the "RCT Data" because it wasn't captured so various stepwise regressions analyzing this variable are conducted purely for exploratory purposes. In order to validate these findings, they would need to be tested out-of-sample on an independent dataset but this was outside the scope of my data collection efforts.

In my testing, I run a series of additional tests in which I weight the number of student-tutor pairs by the square root of the number of sessions for a given pair. For example, previously assume there are two student-tutor pairs, one with an average score of 5 based on 1 session and one with an average score of 4 based on 9 sessions. In my first series of regression, the data enters the regression for ratings as 5 and 4. In my second series of regressions, this would be entered as 5 weighted by sqrt(1) and 4 weighted by sqrt(9) [3]. The purpose of this modification is that student-tutor pairs with more sessions are given more statistical weight in the analyses because the average rating they have is more meaningful as it is derived from a larger number of data-points. This has the importance of putting more statistical importance on longer term student-tutor interactions. This technique is more relevant if the marketplace has a high degree of student-tutor churn after initial matches.

If you consider a network in which 90% of interactions end after a single session but 10% of interactions last for an extended period of time, the information value of the long-lasting interactions are likely to contribute to more meaningful insights into the significance of psychometric sub-traits as the idiosyncratic experience associated with a single hour interaction is reduced and as a result bad ratings or high ratings

that reflect initial optimism about the session, a one-time disconnection, a bad start are down weighted. If a marketplace generally has a high frequency of sticky interactions, it would be less necessary to apply this weighting technique.

Another key consideration is whether I am trying to predict average session rating or individual session ratings. Individual session ratings are likely to have significantly more noise and I would be making the implicit assumption that a pairing with 100 ratings should be given 100x more weight than a pairing with 1 which appears illogical as there would be diminishing returns to marginal data-points on the same student-tutor rating pair.

I use a square root weighting design because variance in the mean scales by 1/sqrt(n) so this is the most logical statistical design of the adjustment to control for variance in session length across pairs.

5.7 Descriptive Statistics

A total of 40,615 ratings were captured in total representing 59.6% of sessions in the data-set (69,779 with some excluded because of variance in the rating system from two versions of the Crimson application used for student-tutor lessons). This is a strong ratings response rate given the completion of ratings is purely optional, has a time cost for the student with a clear but slightly unclear direct benefit for them (ostensibly helping Crimson to "make better matches in the future").

There is an overall cancellation rate of tutoring sessions between a given tutor and a student of 6.3%.

The average rating across this dataset was 4.66 on a 1 - 5 scale.

Appendix Table B.9 shows the minimum, maximum, mean, median, 1st and 3rd quartile across all of the HEXACO's core dimensions, and sub-dimensions for students and tutors.

By assessing the student joining date on the Crimson Education tutoring platform, I found that the average student joined on 26th March 2018 in this data-set, suggesting an average tenure on the platform of approximately ~20 months.

There is a total of 1211 declared female students and 1262 declared male students with 15 declared gender diverse/non-binary. This translates to 48.9% female, 50.5% male and 0.6% gender non-binary. Only 25.8% declared a gender. Gender is an optional field. While the act of filling out gender may result in variance in personality characteristics (compared to those who don't fill out gender), this has no effect on my cross-sectional analysis because I am analyzing the impact of psychometric characteristics on student satisfaction (so any variance in personality by gender is already being captured by the direct measurement of personality characteristics informing the results).

I also report the country of origin of the students given its relevance to generalizability. In terms of country of origin, the most common country was

New Zealand at 4508 students in total. This represented more than double the next closest country of Australia at 2243 students. 2736 were from a country other than New Zealand, Australia, Singapore, Thailand or the USA. The most common countries in the other category are China, Russia, Brazil and the UK.

The average tutor has been on the platform since 8th November 2017 suggesting an average tenure of 26 months (2 years, 2 months) at the date of our analysis.

1197 tutors are female, 1132 tutors are male and 4 are non-binary. This translates to 51.3% of the sample being female, 48.5% of the sample being male and 0.2% of the sample being non-binary.

The most common country tutors were from was New Zealand (2719) with the USA being the next most common (1656) followed by Australia (1503) then Singapore (450) then the United Kingdom (387).

The average rating of a tutor, excluding old ratings, is 4.6 with a median score of 4.9 given the relatively higher proportion of people choosing 5 over 1.

The average age of the students on the platform is 16.2. The average age of tutors on the platform is 22.8.

5.8 Summary of Results

The impact of student HEXACO factors (independent variable) on average rating (dependent variable)

When I ran this regression, factors associated with improved student satisfaction were agreeableness (coefficient: 0.050) and openness to experience (coefficient: 0.058) which were both statistically significant at the 5% significance level (Table 1).

Interpreting these effects, this implies that for a one unit increase in agreeableness in the HEXACO test, the student's rating of their satisfaction is likely to increase by approximately ~0.05. This implies there is a positive correlation between trait

agreeableness and student satisfaction.

A one unit increase in openness to experience in the HEXACO test results in an increase of ~0.06 on the student's satisfaction score. This implies there is a positive correlation between trait openness to experience and student satisfaction.

The regression has a low adjusted R^2 value of 0.01 which suggests only a small portion of the variance in rating can be explained by these factors. This is logical because ~0.05 and ~0.06 are significantly smaller than the 0-5 rating range and the variance in ratings falls within a relatively tight band of a first quartile value of 4.66 and third quartile value of 5. The model has a p-value of the F-Test of 0.00 which is significant at the 5% significance level and implies the model is statistically significant in comparison to a model with a straight line at the mean.

This regression model continues to be statistically significant when tested out-of-sample on the RCT data suggesting these findings are robust. When this step-wise procedure is conducted at the level of the HEXACO sub-dimensions, there is no statistical significance suggesting that the model only works on the 6 core HEXACO dimensions (Table 5.2).

The impact of student HEXACO factors (independent variable) on average rating (dependent variable) weighted by the square root of the number of ratings on each student-tutor pair.

Using square-root weighting, I can see (Table 5.1) that agreeableness continues to be statistically significant at the 5% significance level but openness to experience becomes significant at the 1% level.

When this is tested out of sample on the RCT data, only openness to experience continues to be statistically significant (agreeableness loses its statistical significance). I primarily focus on effect size as opposed to statistical significance given the large number of tests conducted.

At the level of sub-dimensions (Table 5.2), sincerity is now statistically significant across both data-sets and positively correlated with student ratings.

Table 5.1: Student HEXACO Factors and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Agreeableness	0.050**	0.058**	0.042**	0.026
	(0.025)	(0.023)	(0.019)	(0.017)
Openness to Experience	0.058**	0.063***	0.048**	0.037**
	(0.024)	(0.022)	(0.020)	(0.019)
Constant	4.244***	4.249***	4.371***	4.493***
	(0.110)	(0.100)	(0.088)	(0.081)
Observations	1,185	1,185	1,972	1,972
R^2	0.009	0.014	0.006	0.003
Adjusted R^2	0.007	0.012	0.005	0.002
F Statistic	5.437***	8.207***	5.679***	3.365**

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

Table 5.2: Student HEXACO Facets and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Dependence	−0.049***	−0.056***	−0.014	−0.004
	(0.017)	(0.015)	(0.014)	(0.013)
Greed Avoidance	−0.058***	−0.059***	−0.014	−0.019*
	(0.017)	(0.015)	(0.013)	(0.012)
Inquisitiveness	0.062***	0.063***	0.016	0.014
	(0.017)	(0.016)	(0.014)	(0.013)
Liveliness	0.064***	0.065***	0.035**	0.028*
	(0.019)	(0.018)	(0.016)	(0.015)
Sentimentality	0.077***	0.064***	−0.012	−0.020
	(0.018)	(0.017)	(0.015)	(0.014)
Sincerity	0.071***	0.066***	0.021	0.030**
	(0.018)	(0.016)	(0.014)	(0.013)
Social Boldness	−0.041**	−0.036**	0.009	0.009
	(0.019)	(0.017)	(0.015)	(0.014)
Constant	4.112***	4.219***	4.512***	4.565***
	(0.124)	(0.111)	(0.096)	(0.088)
Observations	1,185	1,185	1,972	1,972
R^2	0.053	0.062	0.008	0.009
Adjusted R^2	0.048	0.056	0.004	0.005
F Statistic	9.444***	11.067***	2.193**	2.425**

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

The impact of tutor HEXACO factors (independent variable) on average rating (dependent variable) with the exclusion of the old rating system.

Now I move to consider the impact of tutor personality ratings on student session ratings. After the stepwise regression procedure, I regressed tutor honesty-humility against average session rating (Table 5.3). Honesty and humility is significant at the 5% significance level and the p-value of the F-statistic is 0.02 which is statistically significant. The adjusted R-squared is 0.00 which is very low.

When this model is tested on the second RCT sample, there is no longer any statistical correlation between honesty-humility and session rating. Similarly, none of the sub-dimensions appear to be statistically significant when tested out of sample (Table 5.4).

Unsurprisingly, after weighing by the square root of the number of ratings, no statistically significant model is found (Table 5.3 and Table 5.4).

Table 5.3: Tutor HEXACO Factors and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Honesty Humility	0.046**	0.035*	0.017	0.021
	(0.019)	(0.018)	(0.019)	(0.017)
Constant	4.442***	4.519***	4.606***	4.627***
	(0.067)	(0.064)	(0.068)	(0.061)
Observations	2,177	2,177	1,970	1,970
R^2	0.003	0.002	0.0004	0.001
Adjusted R^2	0.002	0.001	−0.0001	0.0002
F Statistic	5.834**	3.782*	0.791	1.483
Note:				*p<0.1; **p<0.05; ***p<0.01
				Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted

Model

Table 5.4: Tutor HEXACO Factors and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Sincerity	0.036***	0.027**	−0.018	−0.001
	(0.013)	(0.012)	(0.014)	(0.012)
Constant	4.478***	4.550***	4.727***	4.703***
	(0.047)	(0.044)	(0.047)	(0.043)
Observations	2,177	2,177	1,970	1,970
R^2	0.003	0.002	0.001	0.00000
Adjusted R^2	0.003	0.002	0.0004	−0.001
F Statistic	7.350***	4.584**	1.803	0.005

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

The impact of the difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable) with the exclusion of the old rating system.

No statistically significant effect is found using the relative difference between student and tutor personality characteristics on average rating (Table 5.5). Initially, in the first sample, openness to experience appears to be statistically significant but this is no longer the case in the out of sample test on the RCT data. Additionally, inquisitiveness and social boldness appears to be statistically significant but this is no longer the case in the out of sample test on the RCT data (Table 5.6).

After weighing by the square root of the number of ratings, the relative difference

between inquisitiveness (a HEXACO sub-dimension) is positively correlated with average session score (Table 5.6).

Table 5.5: (Student - Tutor) HEXACO Factors and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Openness to Experience	0.050***	0.050***	0.011	0.010
	(0.019)	(0.017)	(0.015)	(0.012)
Constant	4.624***	4.671***	4.670***	4.740***
	(0.016)	(0.015)	(0.012)	(0.010)
Observations	1,036	1,036	1,970	1,970
R^2	0.007	0.008	0.0003	0.0004
Adjusted R^2	0.006	0.007	−0.0002	−0.0001
F Statistic	7.210***	8.616***	0.521	0.784

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

Table 5.6: (Student - Tutor) HEXACO Facets and Average Session Rating

	Dependent variable:			
	Cross Sectional Average Rating		RCT Average Rating	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Inquisitiveness	0.045***	0.040***	0.017*	0.021**
	(0.014)	(0.012)	(0.010)	(0.008)
Social Boldness	−0.023*		0.017*	
	(0.013)		(0.010)	
Constant	4.620***	4.672***	4.681***	4.747***
	(0.016)	(0.015)	(0.012)	(0.010)
Observations	1,036	1,036	1,970	1,970
R^2	0.011	0.010	0.003	0.003
Adjusted R^2	0.009	0.009	0.002	0.003
F Statistic	5.914***	10.220***	3.367**	6.142**

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

The impact of the absolute difference between student and tutor HEXACO factors (independent variable) on average rating (dependent variable) with the exclusion of the old rating system.

No statistically significant relationship is found considering the absolute difference in HEXACO and session ratings across any of the samples (Appendix Table B.11 and Table B.12).

The impact of cancellation rate (independent variable) on average session rating (dependent variable).

No relationship is found between cancellation rates and average session rating (Appendix Table B.13).

The impact of student HEXACO (independent variable) on general cancellation rates (dependent variable)

Student Extraversion and Honesty-Humility are positively correlated at the 5% significance level with general cancellation rates of session. Student Emotionality and Openness to Experience are negatively correlated at the 5% significance level (Table 5.7).

Firstly, I analyze the relationship with emotionality and find a negative correlation between emotionality and cancellation rates.

Next, I analyze extraversion. A positive correlation exists between the student HEXACO trait of extraversion and the general cancellation rate.

A positive correlation exists between the trait for honesty and humility in students and the general cancellation rate of the student-tutor pair.

A negative correlation exists between openness to experience and general cancellation rate.

The impact of student HEXACO (independent variable) on tutor cancellation rates (dependent variable)

No relationship is found between student HEXACO traits and the tutor cancellation rates (Table 5.7).

Table 5.7: Student HEXACO Factors and Cancellation Rate

	Dependent variable:	
	Cancellation Rate b (SE)	Tutor Cancellation Rate b (SE)
Agreeableness		0.032 (0.068)
Conscientiousness		0.048 (0.066)
Emotionality	−0.135*** (0.041)	−0.035 (0.067)
eXtraversion	0.122*** (0.037)	0.090 (0.062)
Honesty Humility	0.078** (0.038)	0.026 (0.064)
Openness to Experience	−0.093** (0.040)	−0.086 (0.065)
Constant	−2.434*** (0.257)	−3.789*** (0.446)
Observations	28,207	28,207
Log Likelihood	−7,500.652	−3,638.955
Akaike Inf. Crit.	15,011.300	7,291.909

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

The impact of tutor HEXACO (independent variable) on general cancellation rate (dependent variable)

This produces statistically significant positive correlations with conscientiousness and negative correlations with agreeableness, emotionality and honesty and humility (Table 5.8).

The impact of tutor HEXACO (independent variable) on tutor cancellation rate (dependent variable)

The findings of this regression (Table 5.8) is quite similar to the previous regression for agreeableness, consciousness and honesty and humility but I note that openness to experience appears in this model. A negative correlation between tutor openness to experience and cancellation rate may occur because those tutors feel less intellectual curiosity and so would be less inclined to want to teach for the intrinsic satisfaction of learning and discussion with their student and as a result are more likely to cancel.

Table 5.8: Tutor HEXACO Factors and Cancellation Rate

	Dependent variable:	
	Cancellation Rate b (SE)	Tutor Cancellation Rate b (SE)
Agreeableness	-0.226^{***} (0.042)	-0.236^{***} (0.049)
Conscientiousness	0.274^{***} (0.044)	0.425^{***} (0.057)
Emotionality	-0.100^{***} (0.037)	
Honesty Humility	-0.197^{***} (0.043)	-0.245^{***} (0.055)
Openness to Experience		-0.150^{***} (0.049)
Constant	-2.282^{***} (0.256)	-2.938^{***} (0.290)
Observations	49,012	49,012
Log Likelihood	−9,749.842	−6,755.712
Akaike Inf. Crit.	19,509.690	13,521.420

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

Table 5.9: Tutor HEXACO Facets and Cancellation Rate

	Dependent variable:	
	Cancellation Rate b (SE)	Tutor Cancellation Rate b (SE)
Aesthetic Appreciation	0.004***	0.005***
	(0.001)	(0.001)
Fearfulness		0.005***
		(0.001)
Inquisitiveness	−0.006***	−0.005***
	(0.002)	(0.001)
Modesty	−0.006***	−0.005***
	(0.002)	(0.001)
Organization	0.012***	0.009***
	(0.001)	(0.001)
Patience	−0.005***	
	(0.001)	
Perfectionism	−0.007***	−0.004***
	(0.002)	(0.001)
Sentimentality	−0.010***	−0.009***
	(0.001)	(0.001)
Social Boldness	0.008***	0.006***
	(0.001)	(0.001)
Social Self Esteem	−0.011***	−0.009***
	(0.002)	(0.001)
Unconventionality		−0.006***
		(0.001)
Constant	0.130***	0.082***
	(0.011)	(0.010)
Observations	49,012	49,012
Log Likelihood	5,044.788	16,327.910
Akaike Inf. Crit.	−10,069.580	−32,633.830

Note: *p<0.1; **p<0.05; ***p<0.01

Effect size is unstandardised beta.

5.9 Discussion of Results

Key Findings

The major findings that resulted from my cross-sectional analysis and were replicated in the independent sample was that (1) session ratings (student satisfaction) are positively correlated with the HEXACO dimensions agreeableness and openness to experience at the 5% significance level, (2) session ratings are positively correlated with HEXACO dimension openness to experience when weighted by square-root of the average number of sessions at the 5% significance level, (3) session ratings are positively correlated with HEXACO sub-dimension Sincerity when weighted by square-root of the average number of sessions at the 5% significance level, (4) the relative difference between the HEXACO sub-dimension Inquisitiveness is positively correlated with average session score. The major findings from my cross-sectional analysis regarding cancellation rates (which could not be tested on an independent dataset because of a lack of data) were (5) a positive correlation exists between student HEXACO trait extraversion and honesty-humility in the student and the general cancellation rate and (6) a negative correlation exists between student HEXACO trait openness to new experiences and the general cancellation rate.

Discussion

My finding that session ratings (student satisfaction) are positively correlated with the HEXACO dimensions agreeableness and openness to experience is logically coherent because a student who tends to be kind, sympathetic, cooperative, warm and considerate is more likely to empathize with the tutor and see positivity in the interactions and also penalize the tutor less harshly for any mistakes or misunderstandings.

My findings that session ratings are positively correlated with HEXACO dimen-

sion openness to experience when weighted by square-root of the average number of sessions is logically coherent because online interactions through video-call with a tutor are reasonably different to a traditional brick-and-mortar classroom interaction and students who are more adaptable and flexible are likely to penalize the experience less for it being a change to their status quo. Other possible explanations would include more tolerance to be taught in different learning styles and approaches. One could imagine a student who is used to being taught mathematics in a particular manner such as doing a high frequency of practice tests being frustrated if a teacher shifted to focusing on teaching content through a lecture format if they were accustomed to the first style and had a low openness to experience.

In analyzing my finding that the relative difference between the HEXACO sub-dimension Inquisitiveness is positively correlated with average session score, I assume that the tutor already has natural competency in the subject they are teaching and is also teaching content that is not particularly unique for them. As a result, it could be the case that tutors with particularly high levels of inquisitiveness cannot easily focus on the content they are teaching. This dynamic of a student with higher inquisitiveness may be more optimal as the student is learning new material and trying to build incremental knowledge whereas the tutor is sharing pre-existing knowledge. A highly inquisitive tutor may be less patient when covering repetitive introductory material.

The positive correlation that existed between student HEXACO trait extraversion and honesty-humility in the student could be explained through the technology system itself. All session bookings are supposed to be made early on the Crimson application and cancelled through the application. If the student has a higher degree of honesty, they are more likely to conform to these rules as opposed to cancelling sessions with their tutors through other channels such as Facebook messenger, WhatsApp or email. They are also more likely to follow the guidelines around

early booking of sessions and thus have a greater window before each session in which a cancellation is likely to occur.

A negative correlation was found between openness to experience and general cancellation rate. This is logically coherent because individuals with lower ratings for this trait feel less intellectual curiosity and as a result, are likely to enjoy tutoring and general academics less and thus are more likely to cancel sessions as they replace this time with other more entertaining activities as they seek to maximize utility.

I found no support for the first five initial hypotheses which I was able to test in this chapter. It appears more challenging than initially assumed to use psychometric matching to improve session outcomes systematically because even in my correlational analysis, few statistically significant effects were found. The only successful example of a psychometric match which produced an improvement in session ratings in considering both the tutor and personality's characteristics was on sub-dimension Inquisitiveness. This is interesting to test in further studies to see how robust the relationship and intuitively makes sense when considering how "Inquisitiveness" is defined.

The major strength of my work was in being able to access and analyze a novel data-set that enabled an assessment of the impact of psychometrics characteristics on students. In none of the existing literature is there any large-scale analysis of student outcomes or satisfaction based on psychometric personality characteristics, partially because it requires a fairly novel experimental design or real world use case to generate such a data-set and few tutoring companies conduct research in partnership with universities or collect personality data in the first place. No traditional high school has large scale psychometric testing of both teachers and students. Additionally, traditional high schools do not capture student session ratings after every lesson. The novel characteristics of the online education environment within the dataset provides valuable quantitative insights in a domain in which previously

only case-studies and small sample size studies have been conducted. The other strength of the work was in the relatively high completion rate of the student satisfaction measurement, despite it being optional. The low friction required for the student to complete the session rating coupled with its lack of mandatory completion requirements is likely to reduce bias (when understood through the lens of the high completion rate).

The lack of an independent second data-set to evaluate cancellation rates is a clear limitation of my work. While a number of potentially interesting relationships were found between HEXACO personality traits and cancellation rates, these findings need to be tested on a fresh sample to validate their statistical significance. As demonstrated by a number of the models that were developed on the Pre-RCT data and then tested on the RCT data only in which the associations were no longer significant, some of the findings may have resulted from randomness. It does appear reasonable that HEXACO dimensions like conscientiousness would have a relationship with ratings (although the positive correlation in my analysis may be more indicative of the optional nature of the rating reporting mechanism in the dataset used) but this needs further consideration.

The relatively high completion rate of ratings in the data set at 59.6% is reassuring as the results are more likely to have missing data issues in which the group of people filling out the ratings vary from those that don't in systematic way if the proportion was low. In analyzing the two cohorts, no statistically significant differences were found between the group's demographic data.

In this cross-sectional study, another limitation of my work was that I was unable to assess any causal impact of student and tutor personality matching on student outcomes directly and was only able to conduct various correlational studies thus unmeasured confounding could have remained an issue. To build on the findings in this initial analysis, a randomized controlled trial will enable a consideration of

whether any real causal relationship exists between psychometric characteristics and student outcomes or student satisfaction. A final limitation of my work is that it was hard to base my initial hypotheses on existing academic literature directly using psychometrics in the context of student-teacher relationships so wider studies had to be consulted. While this cross-disciplinary lens is useful, the subsequent lack of validation for these hypotheses suggests that importing logic from adjacent fields has to be done with great scrutiny and sensitivity to the differences.

Psychometric matching between students and tutors appeared to have a promising impact on student outcomes when conducted in a small sample size, qualitative analysis in my systematic review (Packer et Al, 1998) but this is yet to be validated in any large-scale study. In Chapter Six, I will assess the impact of personality matching of students and tutors directly on student outcomes and student satisfaction over a large sample size through a randomized control trial. While the literature that directly uses psychometric matching is sparse, there is widespread literature that studies the role of the student-teacher relationship on student achievement that this cross-sectional analysis builds on. Existing work finds that a reciprocal model exists between teacher acceptance of the student and academic achievement works in both directions and occurs across a wide age spectrum (Tement at al, 2014). Other relevant work includes "Self Reported Personality and School Achievement as a Predictor of Teacher's Perception of Their Students" which found that academic achievement and personality dimensions and traits that related to socialization were correlated with teacher perception of students (Torrubia-Beltri et al, 1999). My finding that differences in specific HEXACO dimensions had a relationship with student outcomes helps to further build on models of the relationship between the student and teacher in the learning process.

In conclusion, I identified a statistically significant impact of the relative difference in inquisitiveness between the student and the teacher on student

satisfaction which gives some weight to my overarching assertion that the relationship between psychometric traits can play a role in the learner experience. I draw a distinction between the value of my finding that student psychometric traits to student session rating scores as this may be reflective of easier or harsher grading but without any change in the actual lesson quality and the finding related to inquisitiveness. Given the inquisitiveness finding did not appear to come directly from a consideration of student psychometric characteristics alone, there is some evidence to suggest it could be the relationship between the student and tutor that is affecting the student's session score. While this is only modest evidence for my initial assertion, it would be worthwhile for more research to consider how psychometric compatibility assessments could be utilized to improve the student and teacher experience in the learning process to either disprove my findings as erroneous or validate them further and test other psychometric tests, matching algorithms between the student and tutors and other learning environments (in-person, small group class matching, online group class matching and other examples).

6

Randomized Control Trial

6.1 Summary

Background

Cognitive matching between students and tutors appeared to have a promising impact on student outcomes when conducted in a small sample size, qualitative analysis in our systematic review (Cognitive Style and Teacher-Student Compatibility, Packer et al, 1998,). Gender matching between students and tutors had mixed results in our systematic review (Morales et al, 2020, Mwanza et al, 2017, Marsh et al, 2008, Cho et al, 2012). In my cross-sectional study, I was unable to assess any causal impact of student and tutor personality matching on student outcomes directly and was only able to conduct various correlational studies. I was also unable to test the impact of gender, even only by correlation, given this data

was not coded. In cross-sectional analysis, the key findings were:

(1) tutor HEXACO traits had a minimal first order impact on average student session ratings,
(2) student HEXACO traits agreeableness and openness to experience were positively correlated with increased student ratings and these findings were robust across two independent data-sets with small to modest effect sizes.

These key findings relied upon correlational data, and thus unmeasured confounding could have remained an issue. The goal of my thesis was to assess factors that can impact student outcomes and student satisfaction in online schooling. While understanding correlational effects was useful, studying causality where possible was important to achieving my stated research objective. I therefore carried out a randomized control trial, pre-registering my hypothesis. In this chapter, I present this randomized control trial which evaluates the causal impact of HEXACO (personality) on student outcomes and student satisfaction. The randomized control trial was regarded as the gold standard of experimental design because it enabled causation to be determined, as opposed to only correlation. I am matching on personality because I hypothesized that a student-tutor relationship with more complementary personalities would improve the productivity of lessons and drive a causal impact on academic achievement. This is the only randomized control trial conducted so far in this area and is an important contribution to the literature because it tests a potential avenue of algorithmic matching between students and tutors that could lead to systematic improvements in student outcomes and student satisfaction. While this quantitative test is novel, one limitation of my randomized control trial design is that I narrowly considered psychometric characteristics and did not incorporate a broader range of factors in the quantitative modelling.

Given the randomized control trial involves proactive intervention as opposed to analysis of historical data, the standard for an ethical trial was higher. The Univer-

sity of Oxford granted the trial ethics approval (reference: SSD/ CUREC1A/BSG_C1A-19-04). The psychometric surveys being administered (HEXACO) was a study of social perception and interaction that doesn't involve any risk of inducing distress or revealing sensitive information about the participant. The intervention poses no ethical concerns as matching by gender has been well studied in the academic literature analyzed in the systematic review earlier. Psychometric matching is a relatively more novel type of intervention but the use of psychometrics is widespread and is well supported. This chapter is one of the most important contributions to the literature in my thesis because it provides causal evidence for the potential impact of psychometric matching on student outcomes and student satisfaction. It is, to my knowledge, the only randomized control trial in the education matching literature so far.

Methods

I conducted a randomized control trial analyzing the impact of matching by HEXACO (a personality assessment) on primary outcomes (1) student outcomes (centred SAT score improvement) and (2) student satisfaction (self-reported student ratings of tutors following each session). The intervention group of students was matched with tutors using a personality matching algorithm and the control group of students was matched with tutors randomly. The sample was collected on a rolling basis over a three month window and the intervention was conducted over a six month window. The population were 15-18 year old students, globally distributed, receiving online one-to-one instruction from primarily university students using the company Crimson Education. I analyzed the data using 2-sample t-tests and hierarchical ordinary least square regressions. This study is registered in the AEA RCT Registry and the unique identifying number is: AEARCTR-0004444.

In my pre-analysis plan, to measure student outcomes, I planned to run 2-sample

t-tests comparing centred SAT score improvement of the trial and intervention group overall, in Mathematics, Writing and Reading (Collegeboard, 2020). I also planned to run ordinary least square regressions with intervention as a binary variable. To measure student satisfaction, I planned to run 2-sample t-tests comparing average session score ratings between the trial and invention group. I also planned to run 2-sample t-tests comparing the impact of gender-matched student-tutor pairs to non-gender-matched student-tutor pairs followed by an ordinary least square regression with gender as a binary variable.

Findings

I included 1,095 students overall (543 students in the control group and 552 students in the intervention group). I included 703 tutors in the trial. No students or tutors withdrew from the trial. There were a total of 2,058 student-tutor pairs in the control group and 2,040 student-tutor pairs in the intervention group. In total, 150 participants in the control group had SAT results for the first diagnostic and of these 25 participants did the final SAT test. This constituted a 20.0% completion rate of both the first and last diagnostic. 147 participants in the treatment group had SAT results for the first diagnostic and of these 21 participants did the final SAT test. This constituted a 16.9% completion rate for both the first and last diagnostic. I found that (1) matching with the HEXACO algorithm led to a statistically significant improvement in SAT writing performance of 23.00 points (moderate effect size) at the 5% significance level and (2) matching with the HEXACO algorithm led to no statistically significant change in SAT Math or SAT reading. I found that (3) gender matching led to a statistically significant increase in average session rating of 0.169 on a 5 point rating scale (moderate effect size) at the 5% significance level. I also found that (4) student outcomes measured by centred SAT score improvement was not correlated with student satisfaction

scores measured by average session rating. In my exploratory analysis, I also found that (5) matching by the HEXACO algorithm (personality matching) led to a statistically significant increase in student session ratings in Western Europe and Latin America (0.200 on a 5 point scale in Western Europe, 0.224 on a 5 point rating scale in Latin America at the 5% significance level) (6) matching by the HEXACO algorithm (personality matching) led to a statistically significant decrease in student session ratings in East and South Asia and Eastern Europe (-0.139 on a 5 point scale in East and South Asia, -0.207 on a 5 point scale in Eastern Europe at the 5% significance level). Finally, in my exploratory analysis, I found (7) gender matching led to a highly statistically significant impact on the first two lessons of a student-tutor pair and then waned in significance as the number of interactions grew (0.337 improvement on a 5 point scale in first 2 lessons at the 5% significance level, 0.0542 improvement on a 5 point scale on 3 or more lessons).

Conclusion

Matching based on personality (HEXACO) can have a positive causal impact on achievement in SAT writing with a centred improvement of 23.00 points which was significant at the 5% confidence level. Given variance in impact of the HEXACO algorithm by geographic range, there may be an opportunity to optimize the impact of psychometric matching through regional coefficients in each matching algorithm to adjust for any cultural differences in students. Matching based on gender had a modest statistically significant improvement on satisfaction. It appeared that gender matching primarily impacted the relationship in the early stages potentially as same-gender pairs had better understanding or comfort levels from the outset (0.337 improvement on a 5 point scale in first 2 lessons at the 5% significance level, 0.0542 improvement on a 5 point scale on 3 or more lessons). The completion rate of student-tutor participants in the intervention for the student outcome measure

was low as completion of additional tests was optional.

6.2 Introduction

Methods (Study Design and Participants)

I conducted a randomized control trial analyzing the impact of matching by HEXACO (a personality assessment) on primary outcomes (1) student outcomes (centred SAT score improvement) and (2) student satisfaction (self-reported student ratings of tutors following each session). The intervention group of students was matched with tutors using a personality matching algorithm and the control group of students was matched with tutors randomly. The sample was collected on a rolling basis over a three month window and the intervention was conducted over a six month window across 2019-2020. I analyzed the data using 2-sample t-tests and hierarchical regressions.

The students are 15-18 year olds who primarily attend private schools or international schools and have the intention of leaving their country to pursue tertiary studies in the United States of America or the United Kingdom upon completion of high school. They generally are highly academic in the top 10% of their school's academic performance. The students (and their parents) had opted into purchasing tuition services from the platform that are one to one in nature and delivered online with the primary goals of increasing academic performance. The students were primarily European, Chinese, Korean and Indian with European being the most common demographic followed by Chinese. The students included both male and female representation in approximately equal proportions. Of the students that registered gender, 220 were male and 232 were female (48.6% were male and 51.4% were female).

The teachers were primarily traditional college-aged students (aged 18 – 22) from

university programs including Harvard, University of Auckland, Yale, Princeton and other comparable institutions. The teachers were generally not the same nationality as their students. They are contractors who typically engaged with two students through the platform. The teachers are not accredited but had gone through a screening program and a training program before beginning their work with students and have applicable child-safety checks in place. They are compensated for their time based on the number of hours they interact with their students for. They generally stay with the same students throughout the course of the student's program except in instances of sickness or other challenging life circumstances. The teachers were primarily European, Chinese, Korea and Indian with European being the most common demographic followed by Chinese. The students included both male and female representation in approximately equal proportions.

Where possible, I have used routine data collection. Both of the outcome variables being measured were routinely administered for all Crimson users by the company.

Procedure

While all participants sat the HEXACO personality assessment, only the intervention group were matched using the HEXACO matching algorithm as follows:

$$
\begin{aligned}
Maximize\ \{\quad & 24.836 * (Flexibility_{student} - Flexibility_{tutor}) \\
& - 22.579 * (Forgivingness_{student} - Forgivingness_{tutor}) \\
& - 16.528 * (Gentleness_{student} - Gentleness_{tutor}) \\
& + 26.633 * (GreedAvoidance_{student} - GreedAvoidance_{tutor}) \\
& + 13.682 * (Inquisitiveness_{student} - Inquisitiveness_{tutor}) \\
& - 23.965 * (Modesty_{student} - Modesty_{tutor}) \\
& + 17.015 * (Organization_{student} - Organization_{tutor}) \\
& + 9.475 * (Sentimentality_{student} - Sentimentality_{tutor}) \\
& - 18.305 * (Sincerity_{student} - Sincerity_{tutor}) \\
& - 14.661 * AbsoluteValue(Patience_{student} - Patience_{tutor}) \quad \}
\end{aligned}
$$

When a student was being matched with a tutor in the intervention group, the algorithm ran on all possible student-tutor pairs and substituted in the value of the student on the relevant personality characteristics and the value of the tutor on the relevant personality characteristics. The possible student-tutor pairs were then ordered from top to bottom and a request was sent to the top student-tutor pair. If the tutor had not replied within a window of time, the request was then sent to the next best student-tutor pair and the process continues until a tutor accepts the selection.

When a student was matched with a tutor in the control group, the possible student-tutor pairs were randomly ordered from top to bottom and a request was sent to the top student-tutor pair. If the tutor did not reply within a window of time, the request was then sent to the next best student-tutor pair and the process continued until a tutor accepted the selection.

6. Randomized Control Trial

The algorithm was generated by using a random forest regression algorithm which is one of the more powerful machine learning algorithms used for classification by constructing a large number of decision trees during training and outputting the mean prediction of each tree. The technique is well used because it corrects for a more simplistic decision tree approach that tends to overfit. I ran the random forest regression on the dataset used for the earlier cross-sectional analysis. The matching algorithm takes the value of a given student's HEXACO dimension or sub-dimension score and compares this to the tutor's HEXACO dimension or sub-dimension score and applies a weighting. The maximization function seeks to maximize the cumulative sum of all the weighted HEXACO relative and absolute difference calculations for each input.

For gender-assignment matching, the general Crimson matching procedure occurred and I compared the outcomes from student-tutor pairs of the same gender to those with differing genders in the control group. For simplification, we only considered male-male, male-female, female-male and female-female gender matches which constituted materially all of the matches in the data-set but I acknowledged that additional gender option analysis would be preferable although impractical in our data-set. I analyzed only the gender matches in the control group to avoid any risk of correlation between the psychometric intervention and gender matching. The gender-matches I analyzed were subsequently entirely randomly assigned.

In the appendix, I provided detail on the descriptive statistics of regional distribution of the population, age, distribution of starting SAT math, writing, reading and overall scores as well as the distribution of HEXACO trait and sub-trait personalities, average session rating and distribution of average session rating by region. I also provided these breakdowns for the participants in the trial who had both first and last SAT score exam results (to calculate centred SAT score improvement) for the outcome measure. Additionally, I provided the average

number of ratings by each student-tutor pair.

Methods (Randomization and Masking)

As a student who has consented to participate in research signs up for Crimson services and is uploaded onto the Crimson app, a random number generator from Microsoft Excel produced a number between 0 and 1. Students with a number between 0 and 0.5 inclusive were put into the control group. All other students were put into the intervention group. Matching was blinded and neither the tutor nor the tutee knew that they had been matched. I was also blinded during the statistical analysis and did not know which group was the control or intervention group for HEXACO matching until after completion of the major statistical assessments. At the conclusion of statistical assessments, a key was provided by the Crimson technology team which revealed which group was the control group and which group was intervention.

Tutor Characteristics According to the Number of Students Taught

An interesting question to consider is whether tutors that have higher loads of students ("star tutors") vary in some persistent way from other tutors. If this was the case, the variance in star tutor prevalence across the groups could skew the results. After analyzing the distribution of student load by tutor, a reasonable cut-off to demarcate these groups is five students. In the sample, 446 tutors have less than five students and 257 tutors have at least five students.

I found that tutors with five or more students are older statistically on average (25.48) compared to (23.69) at the 5% significance level. There was no statistically significant difference in gender between the two groups of tutors at the 5%

significance level. There was no statistically difference in region between the two groups of tutors at the 5% significance level. There was no statistically significant difference in HEXACO dimensions between the two groups of tutors at the 5% significance level. The older age skew of tutors with more load could be because they are more responsible and can be organized enough to handle more students or they potentially need more income as they are less likely to be at university.

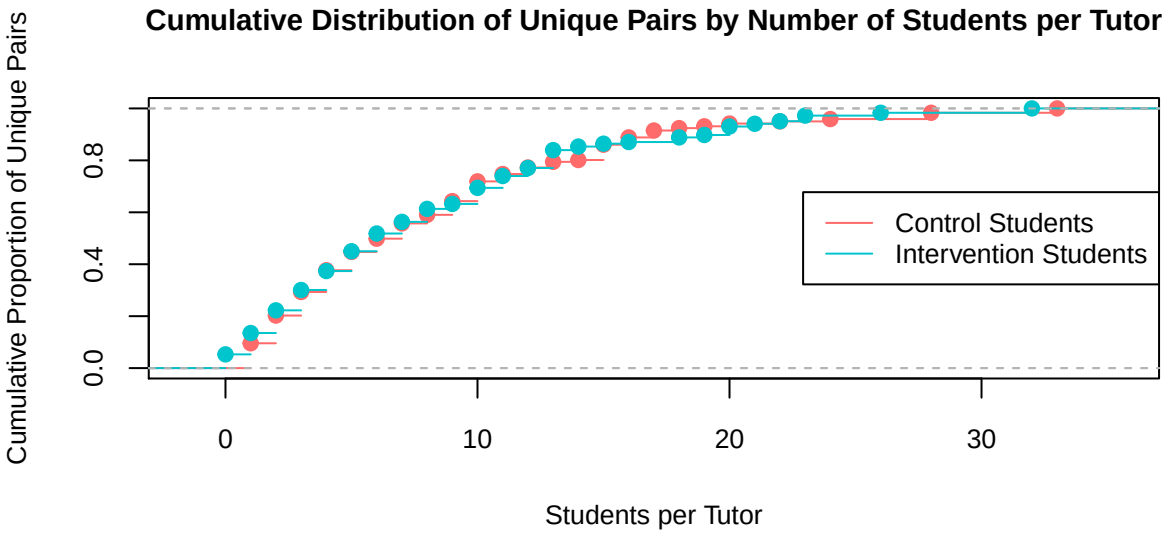


Table 6.1: Welch Two Sample T-Test on Tutors with 5 or More vs Fewer than 5 Students

	Group Mean				
	5 or More	Fewer than 5	Test statistic	P value	df
Tutor Age	25.479	23.693	2.334	0.02	241
Students per Tutor	12.642	1.904	18.644	0.00	260

Outcomes

The primary outcomes were student outcomes (student academic achievement) and student satisfaction.

In order to assess student outcomes, I looked at the improvement in performance on the SAT of students across various diagnostic tests on the Crimson platform.

6. Randomized Control Trial

The SAT is a standardized aptitude assessment. Over 2.2 million high school graduates in the class of 2019 sat the SAT. The SAT assesses three core skills: mathematics, reading and writing and is out of 1,600. The lowest possible score is 400; of the 1,600 points available, 800 is from the mathematics section, 400 is from reading and 400 is from the writing section.

The diagnostic tests on the Crimson platform were designed by analyzing practice exams from the real SAT and adapted by various former examiners and is regarded as a reasonable proxy for real SAT performance. Crimson students generally are required to sit the SAT diagnostic after joining Crimson at least one time and then some proportion go on to sit further diagnostics as they progress towards their SAT. Given only a portion of Crimson students end up applying to American universities, only this portion of Crimson students will typically undergo multiple SAT diagnostic tests. Sitting the follow-up SAT assessment was optional and in a given year around ~15% of students will sit it more than once.

In order to calculate the change in SAT score achievement, we centred each test by subtracting the average score on the given SAT from the score achieved by a student. This helped to correct for any fluctuation in a given SAT test. The tests had been designed to be of comparable difficulty but this adjustment was a prudent way to take an extra precaution so we do not simply pick up fluctuations in the difficulty of diagnostics. This produced the following Change in Centred SAT Score Overall, Change in Centred SAT Math Score, Change in Centered SAT Reading Score, Change in Centered SAT Writing Score. The formula being used can be described as follows:

Change in Centred SAT Score = (Last SAT - Mean(Last SAT)) - (First SAT - Mean(First SAT))

In order to measure student satisfaction, I used a 0 to 5 rating scale. Students were given the option to complete a rating of their tutor after every lesson inside the

Crimson application. Students had to use the application for booking, scheduling and interacting with their tutors. The options were discrete between 0 to 5 and the student could type in a comment. No specific references were provided as to what the numbers mean. If a student did not wish to give a rating, they could click out of the rating tab. This rating is similar to the rating mechanism one is given following the completion of a ride on Uber or a booking on AirBnB. It tends to create a relatively binary distribution with a significant number of ratings being 5 and a smaller subset of ratings being 0 although a reasonable frequency of ratings in the middle exists in the dataset.

The SAT variable had a small number of large outliers which are likely to be caused when a student starts a test and doesn't finish it or answers with random questions. This generated results for one input to the SAT weighting formula that may have arbitrarily low scores.

Outcome and satisfaction data were collected through the Crimson application exported in excel for analysis.

Personality of the tutor and student was measured using HEXACO as described previously in earlier chapters.

Statistical Analysis

To calculate the desired sample size, I estimated a group 1 mean of 4.7 (+- 0.5), group 2 mean of 2.5%, power of 0.8, alpha of 0.05, beta of 2 which implied a suggested sample size of 546 students. This power calculation was based on an assessment of the effect size on student satisfaction scores which was a simpler measure to calculate based on than the significant number of HEXACO traits and subtraits. In total 1,095 students enrolled over a 3 month period to allow for anticipated drop out rates.

I conducted two sample t-tests to evaluate whether there are statistically significant differences in centred SAT scores across the two groups or changes

in average session score. Two sample t-tests were appropriate because both samples were random samples, independently obtained from the sample population. The set up of the randomized control trial enabled samples that were sufficiently comparable for this test to be appropriate given our various balance checks.

I conducted two sample t-tests to evaluate the impact of gender matching on student satisfaction rates.

I conducted ordinary least squares regressions with an independent binary variable which captured whether or not the student of a given pair was in the control group or not and a dependent variable of average student satisfaction.

As students differed in the number of meetings they had with each tutor which could influence student's ratings scores, I ran sensitivity analyses to determine the impact of the number of meetings on outcomes. For each student-tutor pair, I ran three regressions unweighted, weighted by the square-root of the number of ratings for each pair and weighted by the number of ratings of each pair. The unweighted technique doesn't capture any of the additional information that comes from matches with more meetings. One could argue that a nesting technique may be relevant given a number of students may share a given tutor but with ~700 tutors with relatively low numbers of average students per tutor, the effect of any one cluster is relatively insignificant. Additionally, tutors were able to be assigned theoretically to participants in both the trial and intervention group but the tutor was blinded and unaware of how a given student was matched to them so their teaching style was independent of which group the student is in.

I also conducted hierarchical ordinary least square regressions to estimate the effect of student and tutor psychometric characteristics on student satisfaction. The process by which this was conducted was to first run regressions with all HEXACO traits or sub-traits as independent variables against the dependent variable of student satisfaction and then drop independent variables that are not significant

at the 5% significance level. This process was repeated until the final regression formulation only has independent variables significant at the 5% significance level. This does not compare between intervention and control but rather enabled us to understand if student outcomes and student satisfaction are correlated.

For additional exploratory research, I then segmented the data based on region and analyzed the impact of psychometric matching on student satisfaction rates to test for any regional effects. This was not specified in my pre-analysis plan. Additionally, I segmented the gender-matching data based on first two interactions between a student and a tutor and three or more in order to investigate the "first impressions" effect to see whether the impact of gender on a match was consistent over time or had more of an impact initially.

Ethics of the Study

When a student or a teacher (tutor) is enrolled in Crimson, they were given a login to the Crimson App, a learning management system which enabled them to fill out information about themselves including their gender. In both the service agreements for students enrolling and the contractor agreements for teacher enrolling, it was specified that data collected may be used for research purposes which may or may not improve the service experience and that they can opt out of data being used for this purpose by emailing the student's Education Coordinator or the teacher's Tutor Manager. They were then sent a link to HEXACO, the psychometric assessment, with an explanation that the HEXACO assessment will be used to collect information on the person's personality which may be used for the purposes of matching more accurately to students and parents. If the student or the tutor (teacher) chooses to opt out of agreeing to provide their data for research purposes, they were removed from the trial.

These participants were competent youth because they are the decision-maker

that opts into and uses the education services, the research is ethically sound and is not expected to confer any risk of harm, and the psychometric surveys being administered (HEXACO) was a study of social perception and interaction that do not involve any risk of inducing distress or revealing sensitive information about the participant.

The University of Oxford granted the trial ethics approval (reference: SSD/ CUREC1A/BSG_C1A-19-04). The psychometric surveys being administered (HEXACO) is a study of social perception and interaction that doesn't involve any risk of inducing distress or revealing sensitive information about the participant. The intervention posed no ethical concerns as matching by gender had been well studied in the academic literature analyzed in the systematic review earlier. Psychometric matching is a relatively more novel type of intervention but the use of psychometrics is widespread and is well supported.

Conflict of Interest Disclosure: My involvement with the company was a potential conflict of interest for several reasons: 1) I have a large ownership stake in the company, 2) the platform may stand to benefit in some way from the results of the research either directly through improved student outcomes or indirectly through media coverage of any novel findings, 3) the people administering parts of the randomized control trial are employed by Crimson. To reduce these risks, 1) all conflicts were disclosed, 2) I was authorized by the board to publish results, irrespective of their direction, significance or business implication (if any), 3) the publishing of results would put any findings in the public domain, 4) the data cannot be directly accessed by me without making data requests through our technology team which is able to track such requests, 5) I had attempted to find other sources of data which could be used to validate findings obtained through the Crimson platform, 6) I was doing a systematic literature review, as opposed to a general literature review to reduce scope for bias, 7) I analysed and reported on historical

existing data as well as the findings of a randomized control trial which to some degree provided an extra layer of protection as results can be compared, 8) I had used and will continue to exercise strong ethical judgements and continue to be prudent about any conflict of interests throughout the process of the research, 9) the randomized control trial waspre-registered with a pre-analysis plan and any additional exploratory analysis was well documented and 10) I performed the statistical analysis blind to treatment allocation.

Evaluation of Potential Implicit Coercion Risks for Participants

There was no financial penalty or indirect outcome penalty associated with opt out. Additionally, none of the sales team who enrolled students received any commission or compensation for getting students to opt-in to the research and did not have any job outcome or career outcome tied to the proportion of their students, which enrolled in the platform. The same was true for teacher recruitment.

6.3 Results

Balance Checks

In order to test if randomization had been successful, I ran a variety of balance checks. Firstly, the average age of participants in the control group was 16.77 and the treatment group was 16.86. This difference was statistically insignificant at the 5% significance level. 49.6% of participants were allocated to the control group and 50.4% of participants were allocated to the intervention group. I ran a Pearson Chi-Squared test of gender across the two samples and found no statistically significant difference in gender ratios in the two groups at the 5% significance level.

When comparing control and intervention arms, I found no significant differences between the average HEXACO personality characteristics, average starting total SAT scores, math, reading or writing SAT scores of the students.

Table 6.2: Number of Pairs and Students by Student Gender (Control vs Intervention Group)

	Number of Pairs			Number of Students		
	N/A	Female	Male	N/A	Female	Male
Control	1264	415	379	325	112	106
Intervention	1146	439	455	318	120	114

Table 6.3: Pearson's Chi-Squared Tests of Independence on Student Gender (Control vs Intervention Group)

	Test statistic	P value	df
Unique Pairs	13.299	0.001	2
Unique Students	0.569	0.752	2

Table 6.4: Welch Two Sample T-Tests on Student Age, Number of Students per Tutor, Starting SAT Scores and Student HEXACO (Control vs Intervention Group)

	Group Mean				
	Control	Intervention	Test statistic	P value	df
Student Age	16.772	16.859	-0.425	0.671	263
Students per Tutor	2.927	2.902	0.119	0.905	1404
Starting SAT Score					
Overall	19.328	9.505	0.509	0.611	295
Math	10.933	5.618	0.479	0.632	292
Reading	1.950	1.547	0.082	0.935	294
Writing	7.195	-1.232	1.667	0.097	270
Student HEXACO					
Aesthetic Appreication	3.246	3.184	1.201	0.230	1081
Agreeableness	3.072	3.078	-0.188	0.851	1079
Anxiety	3.671	3.555	2.375	0.018	1080
Conscientiousness	3.597	3.606	-0.254	0.800	1081
Creativity	3.574	3.532	0.817	0.414	1081
Dependence	3.001	3.021	-0.361	0.718	1082
Diligence	4.011	3.957	1.277	0.202	1081
Emotionality	3.228	3.207	0.586	0.558	1077
eXtraversion	3.446	3.453	-0.202	0.840	1082
Fairness	3.750	3.726	0.465	0.642	1081
Fearfulness	2.842	2.870	-0.598	0.550	1077
Flexibility	2.940	3.002	-1.337	0.181	1078
Forgivingness	2.843	2.794	1.030	0.303	1082
Gentleness	3.267	3.286	-0.446	0.656	1082
Greed Avoidance	2.824	2.794	0.552	0.581	1071
Honesty Humility	3.328	3.309	0.542	0.588	1077
Inquisitiveness	3.375	3.379	-0.092	0.927	1082
Liveliness	3.470	3.479	-0.200	0.842	1082
Modesty	3.505	3.473	0.690	0.490	1081
Openness to Experience	3.453	3.434	0.597	0.551	1082
Organization	3.341	3.392	-0.947	0.344	1076
Patience	3.232	3.226	0.117	0.907	1081
Perfectionism	3.672	3.669	0.056	0.956	1082
Prudence	3.360	3.401	-0.908	0.364	1077
Sentimentality	3.391	3.380	0.240	0.811	1080
Sincerity	3.229	3.239	-0.206	0.837	1078
Sociability	3.559	3.565	-0.126	0.900	1081
Social Boldness	3.199	3.250	-1.035	0.301	1082
Social Self Esteem	3.550	3.514	0.849	0.396	1082
Unconventionality	3.613	3.635	-0.615	0.539	1081

6. Randomized Control Trial

Given our sample had significant attrition rates in those who sat both the initial SAT assessment and the final SAT assessment, I ran additional balance checks on the group of students that completed both assessments. The gender-balance between control and intervention group was not statistically different between the two groups at the 5% significance level with a p-value of 0.090. The ratio of students from each region was also not statistically different between the two groups at the 5% significance level. The average psychometric match score in the intervention group is -44.54 compared to -59.16 in the control group but this difference was not statistically significant at the 5% significance level (p-value of 0.207).

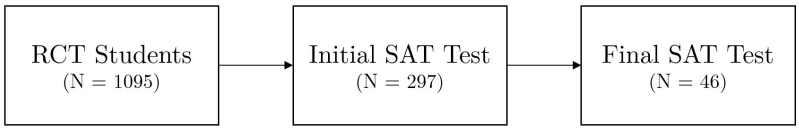


Table 6.5: Welch Two Sample T-Test on Student Age (Across Unique Pairs Containing Students with Valid Starting and Ending Overall SAT Scores)

	Group Mean				
	Control	Intervention	T statistic	P value	df
Student Age	17.197	17.294	-0.211	0.839	6.882

Table 6.6: Chi-Square Test of Independence on Proportion of Students by Gender and Country (Control vs Intervention Group, Across Students with Valid Starting and Ending Overall SAT Scores

	Test statistic	P value	df
Student Gender	6.00	0.199	4
Student Country	23.25	0.277	20

Table 6.7: 2-Sample T-Test on Psychometric Match Score (Control vs Intervention Group, Across Unique Pairs Containing Students with Valid Starting and Ending Overall SAT Scores)

	Group Mean				
	Control	Intervention	T statistic	P value	df
Psychometric Score	-59.156	-44.542	-1.261	0.21	119

How did the students who completed only the initial assessment vary from those who completed both the initial and final academic assessment?

There is a statistically significant difference between the group of students who sat the first SAT and the group of students who sat both the first and last SAT at the 5% significance level. The average age of students who sat the first SAT was 16.52 and the average of students who sat the first and last SAT was 17.26. There was no statistically significant difference in the gender of the students in the two groups. There was no statistically significant difference in the regions the students in the two groups come from. There was no statistically significant difference in starting SAT scores of the students in the two groups.

Table 6.8: Welch Two Sample T-Test on Student Age (Initial Only vs Initial and Final Assessments Completed

	Group Mean				
	Initial & Final	Initial Only	T statistic	P value	df
Student Age	17.259	16.523	2.971	0.0057	31

Table 6.9: Chi-Square Test of Independence on Proportion of Students by Gender and Country (Initial Only vs Initial and Final Assessments Completed)

	Test statistic	P value	df
Student Gender	6	0.199	4
Student Country	50	0.281	45

Table 6.10: Welch Two Sample T-Tests on Centred Student Starting SAT Scores (Initial Only vs Initial and Final Assessments Completed)

	Group Mean				
	Initial & Final	Initial Only	Test Statistic	P Value	df
Starting SAT (Overall)	23.415	12.826	0.423	0.674	66
Starting SAT (Math)	7.752	8.392	-0.044	0.965	64
Starting SAT (Reading)	4.203	1.757	0.356	0.723	62
Starting SAT (Writing)	8.718	1.911	0.931	0.356	58

Matching Check

To check that tutors and students had been successfully matched in the intervention arm, I compared the average psychometric score of student-tutor pairs in the control group and the trial group. I found that the average psychometric score of student-tutor pairs in the control group was statistically higher at the 5% significance level with a p-value of 0.002. This is essential for the validity of the intervention.

Table 6.11: Welch Two Sample T-Test on Psychometric Match Score (Control vs Intervention Group)

	Group Mean				
	Control	Intervention	Test Statistic	P Value	df
Psychometric Score	-44.298	-36.189	-3.147	0.002	2256

Student Outcomes

I found no statistically significant impact of the intervention on centred SAT outcome improvement across total score, math or reading, as shown in Table 6.12. Students in the intervention group significantly improved in SAT writing scores (mean change = 16.44) compared to the control group (mean change = -6.55, p = 0.0039).

Table 6.12: Welch Two-Sample T-Tests on Change in Centred SAT Scores

	Group Mean				
	Control	Intervention	T statistic	P value	df
Change in SAT (Overall)	27.297	16.154	0.404	0.688	42
Change in SAT (Math)	33.862	6.945	1.563	0.126	39
Change in SAT (Reading)	4.998	-3.179	0.675	0.504	39
Change in SAT (Writing)	-6.553	16.443	-2.138	0.039	38

Student Satisfaction

I analyzed the impact of the intervention on session ratings. I found that the control and treatment groups have a virtually identical average session rating score of 4.67 for the control group and 4.66 in the treatment group.

Testing this formally, I ran an ordinary least square regression with an independent binary variable which captured whether or not the student-tutor pair was in the trial or intervention group to see whether the intervention had any statistically significant impact on average session rating.

I found no statistically significant effect of the intervention with a F-statistic of 0.07106 compared to a threshold value of 3.846 with first degree of significance of 1 and second degree of significance of 1,970. The t-test of the F-statistic was clearly not statistically significant at the 5% significance level.

Table 6.13: OLS Regression: Average Session Rating (Across Unique Pairs) by RCT Group

	Dependent variable:	
	Average Rating	
	(1)	(2)
RCT Intervention Group	−0.006	−0.010
	(0.022)	(0.020)
Constant	4.669***	4.705***
	(0.015)	(0.014)
Observations	1,972	1,972
R^2	0.00004	0.0001
Adjusted R^2	−0.0005	−0.0004
Residual Std. Error (df = 1970)	0.482	0.619
F Statistic (df = 1; 1970)	0.071	0.251
Note:	*p<0.1; **p<0.05; ***p<0.01	

Models: (1) Unweighted model (2) Weighted by square root of number of ratings for each pair

For exploratory analysis, I then moved to analyze the impact of the intervention on average session rating by region. This helped to address differences in ethnicity by students with region being an imperfect proxy for ethnicity. It also reflected different parenting attitudes, expectations around educational achievement and education systems.

I conducted 2 sample t-tests to see if there were any statistically significant differences between the student satisfaction scores by region of the intervention. The analysis found the intervention had a statistically significant impact at the 5% level in East/South-East/South Asia, Eastern Europe & Central Asia, Latin America and Western Europe and it had no statistically significant effect in Australia, Middle East & Africa, New Zealand, North America and Singapore.

Analyzing these findings further, I ran ordinary least square regressions to capture the impact of the intervention.

I found the intervention had a statistically significant negative impact on session rating in East/South-East/South Asia with a coefficient of -0.139 significant at the 5% significance level. I found the intervention had a statistically significant negative impact on Eastern Europe & Central Asia with a negative coefficient of -0.207 significant at the 5% significance level. Latin America had a positive coefficient of 0.225 at the 5% significance level and Western Europe had a negative coefficient of 0.200 at the 5% significance level.

To explore these findings further, I cohorted by broad student region and ran t-tests and regressions as follows. I found the impact of the intervention to be significant at the 5% significance level with a negative coefficient of -0.178 across the Asia/Eastern Europe region and to be significant at the 5% significance level with a positive coefficient of 0.223 across the Western Europe/Latin American Region at the 5% significance level. The intervention had no impact on the Other region which includes Australia, Middle East & Africa, New Zealand, North America and Singapore.

In summary, I found our intervention which used psychometric matching of HEXACO generated a statistically significant positive impact on SAT writing and led to higher satisfaction scores across Asia and Eastern Europe and led to lower satisfaction scores in Western Europe and Latin America.

Table 6.14: Welch Two Sample T-Tests on Average Session Rating (Across Unique Pairs) by Student Country

	Group Mean		Two-Sample T-Tests		Weighted Regression	
	Control	Intervention	P value	T stat	P value	T stat
Student Country						
Australia	4.687	4.624	0.267	1.113	0.255	-1.140
East / South-East / South Asia	4.703	4.564	0.045	2.020	0.041	-2.052
Eastern Europe & Central Asia	4.824	4.617	0.001	3.506	0.000	-3.600
Latin America	4.600	4.825	0.020	-2.411	0.010	2.627

Middle East & Africa	4.651	4.651	0.995	-0.006	0.995	0.006
New Zealand	4.728	4.675	0.337	0.963	0.342	-0.953
North America	4.596	4.680	0.273	-1.100	0.267	1.112
Singapore	4.586	4.709	0.105	-1.631	0.111	1.605
Western Europe	4.536	4.736	0.003	-2.960	0.005	2.815
Broad Student Region						
Asia / Eastern Europe	4.772	4.594	0.000	3.971	0.000	4.055
Western Europe / LatAm	4.549	4.772	0.000	-4.135	0.000	-4.063
Other	4.662	4.664	0.938	-0.077	-0.938	-0.077

Gender Matching

I ran an ordinary least square regression with a binary variable capturing gender match. I ran three regressions, unweighted, weighted by the square-root of the number of ratings for each pair and weighted by the number of ratings of each pair. Gender matching did not occur randomly but rather this analysis was conducted on the dataset from the randomized control trial. It may be possible that the lack of systematic randomized gender allocations could create a bias, for example, there may be a greater likelihood of tutors accepting same or different gender matches.

In the unweighted regression, gender matching had a statistically significant effect at the 5% significance level with a coefficient of 0.169 and an F-statistic of 6.936. Both the weighted by square-root and weighted by number of rating regressions were not statistically significant.

Based on these findings, I ran some exploratory analysis to test if gender matching was producing a "first impressions" effect i.e. it had a higher impact on pairs with a lower average number of meetings as the earlier regressions would imply. I split the dataset into student-tutor pairs with two or fewer meetings and three or more. I then run an ordinary least square regression with binary variable gender match. This variable was statistically significant at the 5% significance level with a

large effect size of 0.337 with 92 observations when looking at the cut of data with two or fewer meetings. The variable for three or more ratings was not statistically significant at the 5% significance level with 117 observations.

Interestingly, no statistically significant difference was observed between the gender distribution of the star tutors compared to the overall gender distribution.

Table 6.15: OLS Regression: Average Session Rating by Gender Match

	Dependent variable:				
	Average Rating				
	(1)	(2)	(3)	(4)	(5)
Gender Match	0.169***	0.084	0.007	0.337**	−0.032
	(0.064)	(0.057)	(0.051)	(0.128)	(0.070)
Constant	4.538***	4.634***	4.721***	4.333***	4.719***
	(0.047)	(0.041)	(0.036)	(0.092)	(0.051)
Observations	237	237	237	92	117
R^2	0.029	0.009	0.0001	0.071	0.002
Adjusted R^2	0.025	0.005	−0.004	0.061	−0.007
Residual Std. Error	0.494	0.634	0.903	0.616	0.379
F Statistic	6.936***	2.140	0.020	6.878**	0.205
Note:					*p<0.1; **p<0.05; ***p<0.01

Models: (1) Unweighted model (2) Weighted by square root of number of ratings for each pair (3) Weighted by number of ratings for each pair (4) Two or Fewer Ratings (5) Three or More Ratings

6.4 Discussion of Results

This randomized controlled trial sought to determine whether matching students and tutors based on their personality improved students' SAT scores and student satisfaction scores. I found no evidence that matching on personality improved students' general SAT scores or their scores in math or reading. I did, however, observe a significant improvement in student's SAT writing scores in the intervention arm. I found no evidence that matching on personality influenced student satisfaction scores.

The key findings from my pre-analysis plan were that (1) matching by the HEXACO algorithm led to a statistically significant improvement in SAT writing performance but not on reading or mathematics, (2) gender matching led to a statistically significant increase in average session rating. Other key findings were that (3) student outcomes measured by centred SAT score improvement was not correlated with student satisfaction scores measured by average session rating. Some caution should be exercised in interpreting the SAT findings given a slightly higher dropout rate in the intervention arm, suggesting possible negative effects (16.9% completion in the intervention group compared to 20.0% completion in the trial group).

The key findings from our exploratory analysis was that (3) gender matching led to a statistically significant impact on the first two lessons of a student-tutor pair and then waned in significance as the number of interactions grew. I also found that (4) matching by the HEXACO algorithm (personality matching) led to a statistically significant increase in student session ratings in Western Europe and Latin America, and (5) matching by the HEXACO algorithm (personality matching) lead to a statistically significant decrease in student session ratings in Asia and Eastern Europe.

This study is the first to use a randomised design to test the causal effect of

personality matching on student outcomes, and provides reasonable evidence that further exploration is needed of the potential for psychometric matching to aid in enhancing the quality of interactions between students and tutors. Our randomized control trial, using a personality matching algorithm of certain coefficients, produced statistically significant evidence of improved session ratings (student satisfaction) across Western Europe and Latin America. This has to be viewed through the lens of exploratory analysis as this was not a core hypothesis stated in the pre-analysis plan. It should also be noted that the personality matching algorithm's negative impact on session rating in Asia and Eastern Europe means one must be careful utilizing psychometric matching as certain coefficients could actually harm outcome variables.

Because of the experimental design, this relationship is a causal relationship which means that the psychometric matching was the driving factor behind the improvement in average session ratings in cohorts. Our randomized control trial found a causal relationship between implementation of the personality matching algorithm and an increase in SAT writing.

I hypothesize that the mechanism for improvement in SAT writing is that the student and tutor worked more effectively together because of enhanced personality compatibility translating to improvement in student outcomes. Writing (alongside potentially reading) has more interpretation, ambiguity, emotion and various approaches and as a result would seem more likely to be impacted by a more compatible student-tutor pair than an empirical subject like mathematics with a defined list of concepts, methods and clear answers (Graham, Paz et al, 2002). This could be because personality matching means perhaps the students feel more open to writing emotively and are more able to write freely. I suspect the use of psychometric matching would most enhance humanities subjects that involve ambiguity, debate, emotion and discussion such as art, classics, english, art history, geography, history, languages, media studies, business studies. I suspect psychometric matching would

have relatively less impact on fields like mathematics, computer science, physics, chemistry and biology which are likely to have relatively less ambiguity.

The findings for SAT writing had a meaningful effect size and represent around a 23 point improvement in the SAT writing. Additionally, because the SAT is entirely multi-choice and all SAT answers are graded automatically, there was no room for tutor bias in the subjective assessment of this work.

Given a negative impact was observed on our HEXACO matching algorithm for Asia and Eastern Europe but a wide variety of various traits and sub-traits were statistically significant across different cohorts of analysis, it is possible that a matching algorithm using different coefficients could beat random matching in other countries as well. Additionally, it is probable that the matching algorithm used can be enhanced through consideration of a combination of differences between student and tutor traits and learning from the various regressions. Regression techniques such as random forest algorithms can be deployed to further optimize the coefficients. Alternatively, it may be that our exploratory findings are due to chance. Replication in independent samples is warranted, particularly given the mixed findings across different counties in our sample. Any intervention that has evidence of a potential harm should be taken seriously and the findings across Asia and Eastern Europe mean further research in this area needs to be carefully implemented (Joy, Kolb et Al, 2009).

While it is difficult to ascertain why I didn't find any difference in student satisfaction scores across the trial and intervention group, in general, I hypothesize that the impact of personality matching on student satisfaction varies by cultural background systematically. This is supported by *Cross Cultural Differences in Online Learning Motivation* (Lim et Al, 2007). As a result, across the global data-set used, no universal findings are easily extractable as different regional biases may be working in antagonistic directions to produce no overall findings. It

could also be the case that personality matching simply had no effect on student satisfactions. Another interpretation given my findings around SAT writing score improvement could be that it had no impact on student satisfaction (and student satisfaction doesn't matter for outcomes) but it does produce different final outcome variables. The final impact on academic performance was more important than a proxy of student satisfaction, especially given I found no correlation between student satisfaction and student outcomes in the dataset.

Additionally, the analysis conducted only considers the case of one on one matching. A possible extension of my research would be to one to many applications, most closely a teacher with a group of students (Bloom et Al, 1984). While it is difficult to capture a high frequency of session ratings in a traditional school environment, the growing incidence of online high schools that seek to more rigorously track the student and tutor experience may provide ample data for these types of analysis. While satisfaction data may be hard to come across in the context of brick-and-mortar classrooms other proxies like truancy levels, homework completion rates or effort grades could be used (Conte, Buscha et Al, 2013). Public school chains often have extensive outcome tracking through the use of various diagnostics that are often legally mandated. If large-scale HEXACO psychometric testing is administered, large data-sets could be generated that could be used to build matching algorithms that could be applied to optimize classrooms. Many schools are not sufficiently large in the context of brick-and-mortar schools to warrant large-scale optimization problems because timetabling constraints often are a sufficiently powerful constraint to limit the number of options for teachers a given class may have. Large school chains like Dulwich International, GEMS, Nord Anglia, Inspired Education and the rapidly growing community of online high schools including K12, Laurel Springs, Stanford Online High School, George Washington High School are likely to get to sufficiently large volumes of students to be able to successfully deploy large scale matching algorithm initiatives. While it

would appear unlikely that large group classes can be meaningfully optimized given the large number of interactions that occur between students beyond the student and teacher interaction that form a part of the learning experience, small group classes may be more fertile grounds for testing of this approach. Additionally, an initial batch of models that ignore the social interactions between students and only seek to maximize the compatibility of students and teachers in a pair-wise capacity could be deployed and studied (Martinez, Rubia et Al, 2003). Ultimately, I don't know how one-on-one findings would generalize to group dynamics but it would be interesting for further research.

The variance in outcomes across regions (Western Europe and Latin America, Asia and Eastern Europe, Other) may be driven by cultural differences in parenting styles, learning environments and norms established in the school system, cultural differences in students and the perceived acceptability of tutoring (Bornstein et al, 2012). Western Europe and Latin America are countries with the lowest proportion of students from an Asian background in the sample. Asia and Eastern Europe have the highest proportion of students from an Asian background in the sample. Other countries such as Australia, New Zealand and China have moderate representation of students from an Asian background in the sample. Crimson Education's student community has a relatively high proportion of students from an Asian background in excess of 50% of their global student body. Asian countries tend to have a higher level of tutoring hours consumed per week, higher parental expectations around academic attainment and a higher proportion of income spent on education services. Notably, the world's most valuable (TAL Education, Stock Ticker: XRS) and the world's second most valuable education company (New Oriental, Stock Ticker: EDU) are both based in China, almost exclusively serving Chinese students (William Blair Equity Research, 2019). It would seem logical that variance in the student expectations of tutors by region may contribute to differing impacts of psychometric matching.

6. Randomized Control Trial

Some students in traditionally more relaxed countries with a lower proportion of income spent on education, such as Spain, may be more inclined to want to experience academic exploration and intrinsic love of learning through tutoring as compared to countries like Korea, where tutors are almost exclusively judged by the academic attainment they deliver for students measured by absolute improvement in grades (Sorensen et Al, 1994). I would hypothesize that in countries like Korea and China, student outcome improvement is correlated with long-run student satisfaction more powerfully than in more relaxed education environments like Western Europe and Latin America. This will need to be carefully analyzed in future research.

My findings relating to gender matching are significant because they suggest that gender cannot be entirely ignored in consideration of classroom performance. The persistence of single-sex education in many Western developed countries would suggest there is some benefit to the learning environment of being surrounded by members of one's own gender (Hubbard, Datnow et Al, 2005). Anecdotally, parents often refer to the lack of distraction their child will face in an environment given potentially reduced social pressures for heterosexual individuals or in the enhanced comfort their child will face being surrounded by primarily teachers of a similar gender. Reputed New Zealand single-sex schools such as Auckland Grammar tend to attract almost exclusively male teachers to teach a male student body. Schools such as St Cuthberts, Rangi Ruru, Diocesan similarly are single-sex girl schools that tend to attract almost exclusively female teachers to a female student body.

Our exploratory analysis on our randomized control trial data found gender matched pairs to generate statistically significant improvement in session ratings for the first two sessions. This has a variety of potential interpretations. Students may be more comfortable, engaged or perform better when matched with individuals of the same gender (Koomen, Jak et Al, 2012). The social interaction may be less complicated, especially for heterosexual individuals with age gaps that are

not sufficiently large that romantic affections can be entirely discounted. These findings could be strengthened by a qualitative interview of various single-sex principals and co-ed principals to collect their insights from years of experience working in these different types of environments in order to better understand the causal mechanism behind these strong empirical observations (Carrington, Francis et Al, 2008). My findings as to the significance of gender matching are a meaningful contribution to the literature on gender matching in education because most research papers have limited large-scale empirical evidence of the nature that exists in the Crimson Education dataset.

In conclusion, my research has highlighted the need for more randomized control trials in this field. Given the acceleration in online education, there is likely to be an increase in the availability of large-scale data-sets capturing classroom interactions through platforms including Outschool, Khan Academy, Coursera, VIPKids and others. Randomized control trials are crucial for empirically proving causal relationships in educational settings. In my research, I was able to provide evidence that psychometric matching may improve student writing outcomes but found mixed results on student satisfaction, mathematics outcomes or reading outcomes. Researchers could also design psychometric matching algorithms by region and test their efficacy in randomized control trials to see if our findings are spurious or generalizable. My sample, while global, has limitations in generalizability as the students have self-selected into a paid program and are generally applying to the United States for university. Future samples should include students seeking to apply domestically and ideally also completing programs that are not paid in nature.

Educators and policymakers should support a wide-variety of school choices across both single-sex and co-educational options for families as the evidence suggests there may be some positive benefits of gender-matching and ruling out these options as anachronistic is problematic if the primary focus is student satisfaction and

outcomes (Hoxby et Al, 2003). Policymakers in general should seek to improve data collection efforts. They should also encourage schools to put in place pre and post diagnostic assessments to understand the impact of each schooling year on students and then provide this data readily to researchers who can in turn provide insightful research to guide policies.

7

Public Policy Recommendation

This chapter will provide a number of public policy recommendations derived from my systematic reviews and qualitative analysis as well as from my randomized control trial and cross-sectional analysis. The first major recommendation is the need for further support of broad-based evidence-based assessment in education public policy through gold-standard experimental designs such as randomized control trials. While these techniques have become commonplace in academic research (Sullivan, 2011), a greater degree of public policy could be influenced by this type of work. The second major public policy recommendation raised by the thesis is the need to consider proactive online schooling legislation to prepare for the growth in this type of schooling, accelerated by COVID-19. This has arisen from my qualitative research (Chapter Four) and my systematic review of what drives student outcomes and student satisfaction in online schooling (Chapter Two). I

draw on some of the key findings from the online for-profit university space to inform some of these recommendations with the caveat that principal-agent problems are more pervasive in regulation of K-12 education (given the parent is usually the payer, and the student is the learner).

Throughout my thesis, I have used a mixed-methods approach ranging from qualitative analysis when it came to assessing drivers for students to attend online high school to cross-sectional analysis and randomized control trial techniques when the data-set and population in question allowed for such approaches and assessed the existing body of evidence through a systematic review. Some aspects of education cannot easily be assessed through gold-standard techniques like randomized controlled trials because of practical restrictions like the age of participants or the potential negative consequences of the intervention. I advocate for a more careful consideration of the most robust experimental design technique possible for a given question, with a bias towards quantitative studies.

7.1 The Need for Evidence-Based Assessment in Education Policy Making

Some of the worst offenders of making decisions without clear evidence are often public policy makers. Public policy makers often have strong political incentives that drive some of the decisions that are made (Ivey et al, 2018). As an example, relationships between teacher unions and political parties that depend on their support can often produce outcomes that are not necessarily optimized for student outcomes. In this section, I consider the case examples of New Zealand and Australia. I consider examples of teacher compensation, performance management systems for teachers and licensing. While I didn't analyze these in my own experimental work, they stand as examples in which robust randomized controlled trials or at

least more rigorous empirical work can benefit public policy.

In New Zealand, as a result of government legislation, teachers are paid based on tenure as opposed to student satisfaction or results as per the Secondary Teachers' Collective Agreement. Additionally, teachers are paid the same regardless of which subject they teach and their domain of experience (Ministry of Education Website, 2021). This theoretically may be the most effective way to drive student outcomes but that is hard to imagine. In a market based economy, individuals with rarer skills are paid a higher equilibrium price than those with more commonly sought after skills. Many OECD countries have profound shortages in STEM subjects. An incentive system which pays non-STEM teachers the same as STEM teachers when there are growing STEM shortages flies in the face of conventional economics. If computer science teachers are able to impart a skill to young students which enables them to enter the economy and contribute more relative value than another subject, it would make sense for the incentive structure to reward those teachers more.

Furthermore, for many years teachers are paid based on tenure with guaranteed salary reviews tied to how many years of experience they have. Such a system, which used to exist in the legal industry, disincentives ambitious talent from competing (David Campbell, 2021). As a young teacher, if your salary is entirely determined by age and tenure what financial incentive does one have to exert more effort? The common adage is that teachers often choose what they do because of a love of the field, but relying on these altruistic motivations is not going to radically evolve the labor market for teachers and fix structural shortages in STEM fields. A career in teaching doesn't need to be a kind of modern day pilgrimage. If good teachers are paid competitively for effective work and bad teachers are encouraged to leave the profession, the labor market for teachers can work more effectively. Some individuals would be willing to take a pay-cut and sacrifice working in the technology industry to teach computer science in schools, especially late in their careers, but the bigger

this pay-cut is, the less likely school systems are to solve their STEM shortages.

The lack of performance in management in many public school systems is quite extreme. The New South Wales audit office did a review in 2018 and found that only 53 or 0.1% of teachers in the entire state were formally identified as underperforming or unsatisfactory (Centre for Independent Studies, 2021). The same audit found that only 29 teachers were dismissed or resigned as part of formal action taken based on concerns of poor teacher quality. It is highly improbable that 99.9% of NSW teachers are actually performing at a satisfactory level. Evidence-based assessment of teacher performance would represent a radical change from the status quo in a school system like New South Wales.

A model in which assessment tools like the Cambridge Centre for Evaluation and Monitoring is used to conduct rigorous pre-tests of student competency to measure their growth and the potential value-add of the teacher would pave the way for more meritocratic compensation schemes (John Morris, 2021).

Rigorous experiments need to be run into many of the underpinning assumptions of the current schooling system. In New Zealand, teachers need to pass an accreditation program of sorts to teach in a New Zealand school. This has a number of effects. Firstly, it is very hard for international teachers to enter the New Zealand teaching market. Secondly, it is hard for professionals to transition into the teaching profession from other industries. Consider the case of a PhD in Physics from Stanford with substantial experience teaching classes with rave student reviews. If this individual wanted to teach in a New Zealand school today, they would not be able to. Thirdly, it increases the costs of entering the teaching profession both by extending the time required to enter and the financial cost of the requisite training. In aggregate, the requirement for licensing curtails the supply of available teachers. This serves the needs of teacher unions who don't necessarily have strong incentives to boost supply when they can instead seek to maximize the earnings of teachers by

limiting expansion of the teacher workforce. It doesn't serve the needs of students who need more STEM teachers or more choice in who teaches them.

Critics of this argument could argue that the teaching license system stops bad actors from entering the teaching industry. They could argue that it ensures that people who enter the profession are quite committed to doing so because they passed the license.

A question of licensing should not be a political issue. Instead, it should be an empirical issue. Academics can look at natural experiments of differing cities or countries with different licensing requirements and the subsequent effects it has on the supply of quality teachers. Academics can study whether licensing requirements change the incidence of morally dubious actions by teachers such as sexual harassment of minors. Cost benefit analysis can be conducted to assess how strict the accreditation process needs to be.

Public policy leadership that is at least willing to challenge some of the incumbent incentive structures inside public schooling is required if meaningful progress is to be made in advancing student outcomes and making the teaching profession relatively more attractive for highly-skilled workers to enter.

The liberalization of compensation models for teachers is especially interesting to consider within the context of online schooling. Highly-effective teachers, rather than being constrained to the same class sizes as other teachers, could be used to teach students across a city or a country as opposed to only within a given school. The same kinds of economic models that have unfolded in law firms where highly experienced partners are able to scale their efficacy through teams of associates who do a lot of the heavy lifting could unfold in high school. In a liberalized labor market, a highly effective mathematics teacher could deliver seminars to thousands of students and have junior teachers fulfilling roles such as marking of homework, logistics or manning of in-class question and answer functions. Such

a model may enable the most valuable teachers to earn more which may in turn make the teaching profession more attractive to enter in the first place.

7.2 The Need for Public Policy Makers to Proactively Design Legislation for Virtual Schooling Post COVID-19

My analysis so far has encompassed (1) a randomized control trial and cross-sectional analysis of empirical data from a virtual learning company that operates online high schools and online tutoring, (2) qualitative interviews of existing online high school students from North American virtual schools, (3) a systematic review of the drivers behind student outcomes and satisfaction in online schooling and (4) a systematic review of the efficacy of matching techniques (gender, race and psychological factors) between students and teachers to improve learning outcomes.

After considering the various types of analysis conducted in my thesis, I have found weak evidence to suggest psychometric matching can meaningfully improve student outcomes and some evidence to suggest it can improve student outcomes. I have also learned, primarily through my qualitative analysis, some of the major reasons why a student may consider an online high school and for whom this style of learning may benefit. These findings need to be further considered through specific empirical work to test some of these initial discoveries from my thematic analysis.

One of the most profound questions to ask with any new innovation is where to draw the line. I will consider the public policy implications of my findings from my (1) qualitative analysis of online high school students and (2) systematic review into what drives student outcomes and student satisfaction in online schooling. I will recommend a series of regulatory considerations for government agents, private sector entrepreneurs, prospective parents, students and teachers to consider in

the evolving online education landscape.

I will consider the benefits and costs of existing online education, particularly in the US online university sector (University of Phoenix, Devry University, Strayer University) and in areas such as MOOCs (Massive Online Open Courses) offered by some university providers such as CourseEra, Udacity, EdX and Lynda Learning. I will additionally consider some of the regulator initiatives implemented under the Obama administration in the US online university sector to provide some insight into existing efforts to regulate online universities. Obama's administration conducted, to my knowledge, the most far reaching attempt to regulate the sector. I will make frequent reference to New Zealand, Australia and the United States where a significant number of my interviews and case studies have come from but intend the framework to be globally relevant.

These recommendations are also informed by contextual insights from 50+ individuals who are education domain experts that have been interviewed in the process of developing these policy recommendations. Anecdotal interviews with subject matter experts is a deeply flawed research approach by itself but has been used in combination with the main body of my research to add perspective and additional considerations for things such as feasibility of implementation and practical constraints.

I have highlighted in my qualitative analysis a number of key success criteria for online schooling to be successful. These include (1) self-motivation, (2) desire for academic acceleration through extension coursework, (3) flexibility in schedule, (4) a reduction in stress, stigma and competition relative to traditional brick and mortar schools and (5) parental engagement. The most clear weaknesses of online schooling that detract parents would be (1) less spontaneous opportunities for social interaction, (2) limited access to physical school extracurriculars given the lack of facilities and (3) higher barriers for communication with teachers.

7. Public Policy Recommendation

Of the various success criteria necessary for a student to thrive in online schooling it is important to draw a distinction between what should be dealt with at the regulatory level through hard enforcement mechanisms i.e. restrictions on enrollment and what should be dealt with through soft enforcement mechanisms i.e. education of decision makers such as parents in pursuing these options to make better choices. This distinction naturally has some political implications (Fulton et al, 2002) depending on whether a given regulator tends towards a more libetarian belief (Jowett, Sandra, and Mary Baginsky, 1988) that parents are entitled to have more autonomy over education options for their children or a more conservative view that the state should intervene proactively to avoid risky education achievements (White et al, 1987). I will provide some of my proposed considerations and offer spectrums for regulators to evaluate as well as thought experiments benchmarking virtual schooling against other schooling options to help sense-check the regulatory framework consistency of proposed regulation against brick and mortar schools and homeschooling. While regulatory framework consistency across various schooling options i.e. entry requirements for university admission post high school should not vary across pathways, in some cases, a higher burden of proof may be placed on virtual schools given their relative nascency than to traditional schools which, although imperfect, have existed for centuries without profound change (Gee at al, 2004).

A further consideration is whether virtual schooling should be regulated as a strictly inferior offering to brick and mortar schools only used when brick and mortar schooling is not working or whether it should be regulated as one of a menu of options that for some types of students will be strictly superior to other options. The first framework (coined the "The Solution Framework" for ease of reference) would suggest that students should typically start in brick-and-mortar schools and endeavour to stay there unless some difficulty occurs that has to be solved through an alternative delivery channel. It would suggest stakeholders would

need to prove to the regulator that brick-and-mortar schooling within a certain region has some constraint that is real and also needs to be overcome in order to authorize the use of virtual schooling.

The second framework (coined "The Choice Framework" for ease of reference) would necessitate substantially more choice on the student and the parent to make a decision at any time about whether they wish to pursue full-time brick-and-mortar school, full-time virtual school or a blended combination of brick-and-mortar and virtual schooling (Thorne at al, 2003). My systematic review of student outcomes and student satisfaction in online schooling shows that there is limited evidence to suggest virtual schools, when controlling for student quality, underperform or over perform traditional schools in general, although in absolute levels of achievement, they tend to have more underperforming students in America because of the prevalence of credit-recovery students who are only using virtual schooling to re-sit failed qualifications (Stallings et al, 2016). As a result, I recommend applying the second framework that presents virtual schooling as part of a menu of valid learning options.

The first success criteria to be discussed is self-motivation. Governments cannot regulate that individuals be self-motivated (Walker at al, 2006). Self-motivation is generally classified as either intrinsic or extrinsic. Typically, a market-based system provides economic returns for those that pursue relatively difficult endeavours that solve broadly applicable problems, providing extrinsic motivation (Parayil et al, 2005). In the case of schooling, this may take the form of a student who is motivated to perform well in mathematics because it improves his or her odds of gaining admission into a competitive finance program at a university with a strong track record of highly paid graduates. Intrinsic motivation often comes from alignment of a student's passion (Holcomb et al, 2004) with the subject they are pursuing, a feeling of personal growth in achieving new milestones of knowledge and content

acquisition or a desire to be proud of their achievements relative to peers. Existing regulatory frameworks for brick and mortar high schools do not test any kind of intrinsic self-motivation directly but do typically offer minimum achievement thresholds by which an individual passes to the next grade level.

From a regulatory perspective, it could be possible to administer a psychometric test (Hart et al, 2012) or some other assessment that provides a hard cut-off score to test for self-motivation above which individuals become eligible for virtual schools. This appears excessively heavy handed and difficult to conduct in a legitimate capacity given the lack of a decisive test for self-motivation and the wide variance in student's motivation by subject and by age. Another issue with a test for self-motivation is that students may be unmotivated by the existing brick and mortar school environment before switching to a virtual school, potentially through factors such as bullying or uncomfortable social dynamics (Young-Jones, Adena et al, 2015) , and penalizing a student's flexibility in school choice to pursue an option that may remediate these issues is problematic.

It would be more feasible to ensure students enrolled in virtual schools are self-motivated through education of parents and other decision-makers. This could be done through a disclaimer given to parents opting into virtual schools about the relative significance of self-motivation to success in virtual schooling based on existing academic literature. These recommendations, just like recommendations from doctors, do not need to be binding, but could have a positive marginal effect. This is more akin to a nudge in which the disclaimer should hopefully act to "nudge" parents with awareness that their children are ill-suited for this kind of learning away from enrolling (Thaler et al, 2017). Case studies could be provided as part of the disclosure process of various types of students who have succeeded and not succeeded in virtual schools to help contextualize the types of cases that are most common.

Regulators should probe potential and existing online high school operators how

they check for self-motivation of their students. While it is difficult for regulators to impose a uniform global test for student motivation across all schools, it is much more feasible for regulators to approve licenses to high school operators who have demonstrated a clear methodology for vetting student motivation levels. Different potential assessments of self-motivation could include a student application essay or statement of purpose, reference letters from existing teachers or other community stakeholders who can vouch for the motivation levels of the student and an interview process in which questions specifically addressing self-motivation are delivered. A hard proxy for self-motivation is academic grades which, although biased by environmental factors, socioeconomic factors and other covariates (Halawah, Ibtesam, 2006). Some institutions may wish to administer hard academic grade cut-offs for student entry based on an enrollment exam or other assessment which while imperfect would provide some gauge of student motivation levels. The more complicated the interview process or enrollment process with steps that necessitate the student doing work such as application essays, the more powerful the signal of motivation the student is showing to the admissions committee of the high school. The tests for self-motivation may also vary based on the type of virtual school i.e. a credit recovery institution designed to get students to minimum possible levels of achievement compared to an accelerated STEM (science technology engineering mathematics) acceleration school with difficult coursework.

The second success criteria of a desire for academic acceleration in the student is valid but is likely not sufficiently well-defined to be a hard constraint. Requiring all students to pursue virtual schooling only if they academically accelerate in absolute terms relative to their age level would remove the option for some student athletes, students with social anxiety, those pursuing credit recovery and other options. If a regulator wanted to pursue a hard constraint on academic acceleration under "The Solution Framework", a regulatory environment for part-time virtual school enrollment could necessitate that the pathway can only be pursued if the

student's physical school did not offer the desired course the student was pursuing. An additional constraint could be that students can only pursue virtual schooling if they wish to take subjects in addition to the maximum number of subjects their traditional brick and mortar school offers.

The third success criteria for student achievement in virtual schooling of flexibility of schedule does not need to be enforced through a regulatory framework. Virtual schooling, given the lack of need for students to commute to a physical location, the reduced costs associated with travel back and forth, reductions in expenses such as uniform costs, the lower cost of delivery for the school as a result of a reduction in depreciation expenses of physical land assets and a larger supply of potential teachers generally make virtual schooling consistently more flexible for families. Virtual schools in America already compete to offer flexibility to students in the form of year-long course options, holiday learning options, multiple times for key courses, flexibility around students who prefer to work later in the day or earlier in the day and most importantly limited restriction on where students attend class from.

As opposed to worrying about whether virtual schools will offer sufficient flexibility to students, a regulatory body should be more concerned about offering too much flexibility. Online universities such as University of Phoenix came under intense scrutiny from the Obama administration for accepting virtually anybody that applied, regardless of their aptitude for the degree in question (Coutts, Sharona, 2009). Additionally, they often offered limited to no compulsory class attendance and assessment formats that did not necessitate much knowledge of the curriculum being covered. This translated to students spending very little time in the presence of faculty, only working to complete specific assignments requiring only small sections of the curriculum and resulting in charges of being a "degree mill" in which university degrees are issued of questionable or no value. The "degree mill" was catalyzed by easy government funding in which virtually any student could take

out government loans that the University of Phoenix would make people aware of and help them complete to fund their education.

Online schools have a number of strong incentives to not abuse flexibility and become "degree mills". Firstly, reputation is important in education for driving word of mouth referral in the community. A strong reputation of an institution is likely to reduce the cost of acquiring new teachers, improve retention of staff as they are more proud of their institution and surrounded by higher quality peers, attract more students and subsequently more tuition revenue, attract higher quality students which improves the peer-effects learning experience in which students learn from one another and subsequently lead to either higher prestige (often the objection function being optimized by public schools with constrained enrollment), higher profitability (often the objection function being optimized by private sector entrepreneurs) or more scale (Van Vught, Frans, 2008). Secondly, schools that abuse flexibility may lose their license to operate virtual schools (assuming a license is required as it is in New Zealand and US jurisdiction). Thirdly, excessive flexibility that leads to inferior student outcomes will be visible in student achievement data, university matriculation rates and drop-out rates.

A regulatory framework to address the potential to abuse flexibility could necessitate some equivalency with the achievement objectives of traditional schools. Regulators could require that virtual schools require students to at least attend the same number of classes that brick-and-mortar schools require and keep these aspects of the regulation consistent across both learning formats. Additionally, full-time virtual schools could be mandated to have a minimum number of contact hours per week with educators. Regulators should be careful to mandate a minimum number of contact hours per week for academic extension coursework that goes above and beyond the minimum learning load recommended for a given age range. Take the case of a high achieving academic student who is performing well in

classroom subjects at his traditional brick and mortar school but wishes to take a multi-variable calculus class online. The student may be very capable of self-study guided by self-paced asynchronous videos. Mandating the student attend taught online classes given they have opted into an optional subject they didn't need to take is an unnecessary constraint that doesn't sufficiently recognize the student's positive intentions and demonstrated incentive to learn. Part-time virtual schools could have a minimum number of contact hours per week mandated for credit recovery subjects as reducing the chance a student who already failed a class, fails again is more logical than putting limits around successful, accelerated students.

Contact hours provide an accountability mechanism for students to continue to engage with content on a regular basis and avoid the potential of excessive procrastination in which students defer consuming content until the last minute. While some universities do not have any kind of compulsory contact hours, students in high school have often not developed sufficient self-regulation to be very flexible. Additionally, clear evidence in analyzing MOOCs or Massive Online Open Courses suggests the drop-out rates for fully asynchronous video courses exceed 95% and that adding live mentors and human contact greatly improves contact hours (Ghada El Said, 2017). Online schools which are almost entirely synchronous learning such as the Stanford Online High School have very low drop-out rates in courses (<5%) and schools with high proportions of asynchronous learning content such as Florida Virtual tend to have much higher dropout rates so it would appear to do limited harm adding a minimum contact hour constraint on a weekly basis for virtual school students. Synchronous is defined as a delivery style which uses live lessons in which a teacher and a student are actually interacting (such as a video call). Asynchronous refers to the delivery of content to students without direct interaction with a live teacher (often through recorded video lessons or practice questions). Synchronous delivery by definition involves student-teacher contact hours. Asynchronous delivery does not involve student-teacher contact hours.

7. Public Policy Recommendation

The fourth success criteria is a reduction in stress, social stigma and competition relative to traditional brick and mortar schools. It is important to note that a student opting to leave a brick-and-mortar school for reasons associated with social stigma, bullying, intensive competition or other experiences is leaving a single school that may not be representative of all schools. Schools vary enormously in level of academic intensity from Raffles Institution in Singapore, which generates more admits to Oxford and Cambridge in most years than any other high school in the world to Albany Senior High School in New Zealand, which gives students a full day a week to pursue an independent "impact project" which can be on any topic they like. Online schools such as the Stanford Online High School (Raymond Ravaglia, 2007) which has an average ACT score of 34/36, equivalent to an average accepted score of an Ivy League student are likely to have cultures that put strong pride on academic achievement whether others such as Laurel Springs, which primarily is known for catering to athletic students is likely to have a weaker academic culture. The variance between schools within each category across virtual schooling and brick-and-mortar schooling, from my qualitative interviews, appears larger than any systematic gap between virtual schooling and brick-and-mortar schools.

Different approaches can be used to reduce some of the stress associated with competition.

Firstly, governments can recommend grade non-disclosure policies. In many of my interviews, students stressed that having other students ask them their grades following assessments was one of the more psychologically challenging social interactions. In programs like Stanford Business School, students do not disclose their academic grades. They can log in and check their GPA (grade point average) which is a measure of their overall academic achievement and can possibly check their class rank (or how well they are ranked relative to other students) but they cannot directly ask each other. This mechanism provides sufficient confidentiality

for students who wish to opt out of competition or don't feel comfortable sharing their grades while still capturing many of the benefits of relative competition driving everyone to work harder.

Secondly, public policy makers can recommend the use of competency-based learning and streaming systems over age-based progression mechanisms (Barnard et al, 2020). Student feedback and commentary from multiple principals I interviewed suggests that people often find competition motivating when the competition is reasonable because the peers are of similar ability. When students sit in a classroom environment and their peers are vastly stronger academically or vastly weaker academically it can either be depressing or uninspiring. Rather than allowing students to progress based on age, in an online school environment, students can be clumped based on ability within a given subject. This subject based banding is what the Singaporean Ministry of Education is advocating, one of the world's most successful education ministries by academic achievement (Singapore Ministry of Education Website, 2020). Clumping based on ability removes the stress of being benchmarked everyday to students much stronger than yourself but still captures the motivational effects of peer-competition. Additionally, streaming based on ability means that teachers can teach to a more comparable ability level. Tomohiro Hoshi of Stanford Online High School notes that their competency based framework results in substantially more engaged faculty who have a more enjoyable teaching experience and an increase in the academic motivation of students as they are inspired by their peers across a range of their classes.

Streamed classes can result in detrimental motivational implications for students if they are publicly marked as being in a relatively lower stream. A New Zealand Rhodes Scholar, who was in the fifth highest class at Auckland Grammar School, well-known for their broad streaming system, remarked that the streaming system at his school produced a demotivating "Grammar slide" where students continue

to fall, losing progressively more motivation as they become more embarrassed by their falling academic stature and attempt to build more and more of their personal self-worth from sporting or other extra-curricular activities. As a result, it is recommended that streaming take place but potentially without class names that enable students to easily pinpoint where their class stands relative to other classes.

To further reduce stress, it is recommended that grade boundaries are carefully examined to reduce the potential stress associated with marginal variances in testing achievement. Small variances in grade boundary design can have substantial psychological implications. In Cambridge International Examinations, students are given an A* if they achieve 90% or more and in recent years, specific percentages were given. Now, a university can compare between a student with a 95% or a 96% in Biology as opposed to two students with A* in Biology. There is a diminishing marginal return to time spent studying when students are fighting to optimize over a final few marks. Having stressful, transparent methods for students to compete at a very granular level adds potentially unnecessary pressure in an already competitive process.

The fifth success criteria, which is arguably the most important according to my systematic review on student outcomes in virtual schools, is parental engagement. I define parental engagement as the activities that parents conduct to support the child's achievement in school, often by consulting with their teachers, providing a supportive home environment, taking an active interest in the learning outcomes of their children, finding remedial or extension help where necessary and acting as a positive role model for the importance of academic achievement. A literature review by Geoffrey Cox in 2018 argues that parent-student-teacher alignment is one of the most powerful drivers of learning outcomes. Naturally, a child spends more time with their parent than any single teacher and their parent is with them throughout the entire learning journey so sets a consistent tone around how a family

celebrates academic achievement, makes time for study, prioritizes exams, discusses content from school and puts time into learning more generally.

When a child goes to a traditional school, they are surrounded by many peers socially who are all influenced by their collective parents in forming their personal learning norms (Zhang et al, 2016). As a result, the influence of a parent is relatively diminished because of the less concentrated relative exposure as a higher share of time is spent with other children. In virtual schooling and homeschooling, much more of one's education takes place in the presence of a teacher. In homeschooling, the most extreme example, the parent often plays the role of the teacher as well as the provider of pastoral care.

On a spectrum of parent dependency, home schooling is the most dependent on the parent, traditional brick-and-mortar schooling is least dependent on the parent and virtual schooling is in the middle skewed towards brick-and-mortar schooling. Students in virtual schools still have a full suite of teachers and peers in every class but teachers are typically marginally less accessible for everyday questions and the lack of peer interaction often reduces the level of psychological focus on upcoming assignment deadlines or exam periods because their is less exposure to the common stresses running through other students outside of classes. The student during virtual schooling is often learning from home, often in their bedroom, and is likely to spend more time with parents, especially if one is at home during the day during lunch time and other breaks, as opposed to other peers. Some students will take group video calls with peers during break times but it is less common. Simple activities like going to eat food which at a school would necessitate social interaction may lead to sitting in the kitchen with a parent. The compounding effect of this additional exposure to parents and lack of exposure to students means that regulators need to be relatively more careful to consider parent engagement in virtual schools.

7. Public Policy Recommendation

There are naturally a variety of potential regulatory paths forward in addressing parental engagement. From a conservative policy perspective, parents could have to go to an extensive training equivalent to a combination of a teaching and pastoral care course in order to be qualified to be the adult supervision around the child during their virtual learning experience and be personally liable for negligence charges if the student does not attend class and maintain basic education responsibilities. From a libertarian policy perspective, one could argue that nobody, except potentially the child, is more vested in the child's outcomes than the parent and subsequently there is a limited need to regulate the parent's involvement because they are heavily motivated to ensure their child does well. Most regulators have basic child protection laws around leaving children unattended, feeding children and other basic criteria so it is safe to assume that while the majority or even the significant majority of parents have good intentions, it is important to regulate against a small proportion of bad apples that may seek to exploit a more flexible schooling environment.

It is important to draw a line, regardless of one's personal beliefs on the role of parenting in schooling, around the relative risks of virtual schooling so it is appropriately regulated relative to other forms of schooling. Home schooling places a concentrated bet on the quality of the parent as a teacher and pastoral care role model. A home schooling parent who has limited other resources except perhaps some asynchronous video courses or textbooks poses substantially more downside risk to a child's learning outcomes than virtual schooling where at best the parent is still only a source of pastoral care. Regulators who have already formed a view on the conditions necessary for home schooling to take place should typically hold parents engaging in virtual schooling to the same or a weaker standard but not a more excessive standard. It is relevant to note that home schooling is a rapidly growing learning option pursued with more prevalence in New Zealand, Australia, the UK and the USA than previous years with annual growth rates typically in excess of 8% (Tomohiro Hoshi, 2019).

7. Public Policy Recommendation

As stated earlier, the weaknesses of online schooling that typically worries parents from my interviews with students in my qualitative analysis were (1) less spontaneous opportunities for social interaction, (2) limited access to physical school extracurriculars given the lack of facilities and (3) higher barrier for communication with teachers. The question I should ask is what, if anything, should regulators do about these concerns directly?

Firstly, I address social interactions. Virtual schooling, for those who do not like this schooling option, imagine teenagers sitting in their bedrooms devoid of friends, never leaving the house, participating in no activities except the basic minimum necessary academics to get by with no opportunity for serendipitous social interaction. Currently, regulators do not enforce any particular types of social interactions on individuals. While many people perceive virtual schooling students to be anomalies, and by definition of their learning choice, that may be true, my interviewees attending virtual schools were often well-rounded individuals with expansive social networks with peers from their virtual schools and from their communities and other extra-curriculars. In a brick-and-mortar school, if a student wishes to attend no social activities, not speak in class, resist making friends and participate in no extra-curriculars, this is very possible under the existing regulatory framework. Many teenagers spend the majority of their communication time on mobile applications (Goldman Sachs, 2019). Virtual schooling is not likely to be the nail in the proverbial coffin that makes a student swing from being socially well developed to being unable to function. It is not realistic for regulators to mandate any particular extra-curricular activities directly or to require peer-to-peer social interactions as a prerequisite activity for graduation.

Rather than a heavy-handed regulatory approach, I recommend regulators encourage virtual school operators to actively describe how they will foster strong social interactions and student engagement in their school. Often virtual schools have

extensive time for discussion in the synchronous class time interactions providing students a comparable level of peer-to-peer interaction as in an offline class. They typically have meetings such as assemblies and other community activities that bring individuals together for larger meetings to create community norms and a shared culture. Additionally, they often offer meet-up opportunities with teachers and fellow students in certain locations. Stanford Online High School, for example, regularly hosts events on campus so virtual school students can travel there to meet other students or teachers.

It is also recommended that a proper pastoral care infrastructure is demanded of schools where a teacher, who does not teach any of a student's particular subjects, checks in with the student regularly to see how they are progressing. While perhaps too heavy-handed to require, virtual schools should generally have trained mental health resources and other types of counseling support available for students. The potential scale of virtual schools to tap into much larger student communities than physical campuses may make it more economical for operators to have larger staffing cohorts in these areas where typically schools are under-resourced.

Secondly, regulators may be concerned about limited access to physical school extracurriculars and laboratory spaces. Many physical schools do not have extensive or any extra-curricular offerings of substance and lack the necessary facilities to teach laboratory work. Some students may not care at all about extracurricular opportunities at school because they do not pursue any or potentially the ones they do can already be accessed through clubs outside of school or existing coaches. One of the most popular use cases of virtual schooling at Laurel Springs, a for-profit online high school, is for student athletes who play competitive sports that require them to regularly travel to have a schooling option that suits their travel schedule. A competitive golf player doesn't need Laurel Springs to have their own golf team because they can already access superior activities through their club.

Rather than requiring virtual schools to have various facilities which would be clearly impractical given the distributed nature of the students and teachers of a school, regulators should encourage operators to set up communities of shared physical resources and extra-curriculars their students can access. Virtual schools can create partnerships with local libraries, cafes, university laboratories, physical school orchestras or theatre groups. For virtual schools and more generally for physical schools, rather than attempting to provide a jack-of-all trades experience, it is more logical, especially for public operators to partner with existing community assets that are often underutilized and focus resources on improving academic delivery. Students can be provided with lists of local extracurriculars and even be required by the school to participate in a certain number of activities, something that a small selection of private high schools require of their students.

7.3 Principal-Agent Problem in Virtual Schooling

A notable difference between virtual schooling and online university education is that generally in virtual schooling, the parent is the stakeholder paying for the education but the student, often under the age of 18, is the one participating in the education (Ensminger et al, 2001). With online university education, even for bachelor's degrees, the individual paying for the degree is often the one participating in the education (Lane et al, 2018). This introduces a slightly riskier principal-agent problem. In the university system, a student who feels they are not getting value for the money they are paying is more likely to quit the program and find an alternative provider because they are generally internalizing more of the economic cost themselves. With a virtual high school, students are often quite removed from the tuition amounts and what they mean economically to the parents relative

to other expenses. Additionally, they are quite divorced from what good value necessarily is as a high school diploma is part of an investment into building a necessary set of tools to one day day become employable and earn a competitive market salary (Hart et al, 2019). This means that in the event a student is not receiving a rigorous or engaging education, they are relatively less likely to complain, terminate their program or perhaps even be aware of the deficiencies as they may have limited experience of alternatives. Regulators should encourage operators to use hard achievement statistics such as test scores or satisfaction rates that are benchmarked across suitable peers (Net Promoter Scores, for example) to assess how an institution is doing. Additionally, regulators and operators alike should pay careful attention to whether the parent is the stakeholder providing feedback or the student. It may be the case that surveying the parent satisfaction of their child's virtual learning experience compared to the child's satisfaction of their own experience will produce different results. Both perspectives are likely to be useful in receiving a holistic view as parents have more experience across different types of education and consumer experiences and students have first-hand experience with the virtual school platform in focus.

The risk associated with the principal-agent problem in virtual high schools can be reduced by aligning the student and parent experience through parent-teacher communication, giving parents access to recordings of student classroom experiences, student homework, student grades, other parents and regular, healthy communication between the parent and child at home (Miller et al, 2005).

7.4 A Case Study of Regulation in US For-Profit Online Universities: The Good, The Bad and The Ugly

One of the most relevant case studies to analyze for regulators looking to release a regulatory framework to regulate or fund virtual high schools is the US for-profit online university sector. The US for-profit online university sector has made more revenue than any other segment of online education in the history of the world (Morgan Stanley, 2019). For-profit universities have been around since the early 19th century with the first business college (Bartlett's Commercial College) opening in 1834 by Montgomery Bartlett (Kinser, 2006). Historically, college education in the United States was based on a liberal arts curriculum described in the influential Yale Report of 1828 as a classical curriculum that fostered sufficient mental discipline to be a respected member of society. As individuals demanded a more vocationally oriented college education, many for-profit institutions continued to expand rapidly with more than 200 enrolling more than 80,000 students by 1850 (Kinser, 2006). As of 2019, although only 5% of Bachelor's Degrees are granted at for-profit universities, 12% of all Bachelor's degrees in Business, Management and Marketing are granted by for-profit universities. As the for-profit sector grew, the companies became increasingly commercial, tapping into public funding through initial public offerings. Companies such as Apollo Education, the parent company of University of Phoenix listed in 1994. Other companies such as Devry Education, Strayer Education and Capella University trade in 2019 and are accountable to Wall Street shareholders that assess their financial performance, including student enrollment every quarter. These companies enjoyed substantial growth and between 1998 and 2008, enrollment at for-profit universities grew by 225% as sector enrollment grew from 766,000 to 2,400,000 students (Lee, 2012).

7. Public Policy Recommendation

For-profit online universities have several analogies to virtual high schools. Both types of institutions generally have flexible curriculums that enable learners to progress through their qualification at a pace that suits their learning ability. This often leads to individuals completing coursework faster than recommended completion periods if they are high achieving or slower than recommended completion times if they are relatively underperforming academically. I have referred to this as "competency based learning". Additionally, both types of institutions deliver their content through a combination of synchronous video calling and asynchronous lecture content. Instructors are also available through office hours and students often sit assessments online or participate in group projects with other students. Most students complete these degrees from remote locations, typically at their home. Both types of institutions typically cost less than private institutions of the same nature often because of operating efficiencies relating to the lack of physical assets and leases. Both types of institutions typically cost more than public institutions which are entirely funded by the government. For public virtual high schools, the cost to the government per student is typically lower than the cost to the government per student of brick-and-mortar education for similar reasons. In America, both types of institutions also tend to be subsidized by the government in some form either through accessible loans or direct tuition payments. In both types of institutions, students can participate in classes consisting of diverse learners from different parts of the state or country.

From a public policy regulatory perspective, the concerns around for-profit operators in both spaces are quite similar. Are the operators enrolling students who are likely to be successful in these institutions? Is the marketing and sales strategies used by these institutions accurately reflecting the learning experience and graduation outcomes such that prospective students can make an informed choice? Is the price-to-value ratio fair based on what the institution charges for tuition and what the student receives? Is the institution being used by students

as a way of skirting academic standards and pursuing a more easy pathway to a diploma than the government intends? Are the graduating outcomes at the institution comparable or better than peer institutions for students of the same background? If the government is providing funding to the institution, is the return on investment of this money beneficial for society relative to the set of other education initiatives that could be pursued?

7.5 Learning from For-Profit Online University Regulation

Regulators should learn from the multi-decade development of regulation around the US for-profit online university sector (Ann Morey, 2004) to regulate around the most problematic areas and reduce the risk of student exploitation while maximizing access and the learning benefits of the delivery method. I will outline three of the key learnings from the regulation that are worthwhile to understand. Firstly, the 90/10 rule, which describes the ratio of revenue an institution can receive from the government relative to private paying individuals. Secondly, the gainful employment, job placement and default rate regulations that track the performance of institutions in preparing students for success in the economy. Thirdly, the funding model of government investment in the sector to drive more equitable access to the benefits of online education.

Firstly, the 90/10 rule restricts the amount of revenue that for-profit online universities can receive from governments to 90% of total revenue from Title XI funding. Title XI funding is Federal Student Aid often in the form of loans. An average for-profit online university receives 71.5% of funding from Title XI funding and if they exceed 90% they may lose eligibility for this funding which would be a crippling blow for an institution because at this point this would result in a 90%

reduction in revenue (Matthew Hodgman, 2018). The constraint is designed to be so painful for any institution that universities stay comfortably within the range. The logic behind this regulation is that it is relatively easier to convince an individual student to take up a loan to pursue an education. Both the cost of the loan and the benefits of the education are going to be experienced in the future so a student is likely to be more willing to enter into such an agreement than an individual who has to pay with their own money upfront. Ensuring an institution has a sufficiently large body of private pay students is a health check on the price-value ratio of the virtual school. One could imagine an institution that receives all of its revenue from the government that charges an excessive price that is above what a normal, private paying student would pay because individuals are often not very effective at appropriately assessing the total cost of a loan over the life of the payments (Kahneman and Tversky, 2011). Private paying students who are forking over hard earned cash every semester are more acutely aware of what they are receiving, what they are paying for and the economic trade-offs they are making in their life at the moment to facilitate their education. The other rationale behind the 90/10 regulation is to avoid universities preying on low-income students who are unlikely to be able to afford the full debt burden of their education. Typically, Title XI funding is income based so universities that receive all of their funding from this channel will have students from a disproportionately low socioeconomic background.

Regulators should consider implementing a maximum cap on the proportion of tuition revenue a virtual high school can receive from public funding. Ministries of Education are more concerned about ensuring universal access to high school education which is viewed as a right than universal access to university education, so some governments may not want to impose any constraint on the source of tuition revenue. In this case, they should carefully track other indicators of student outcomes to ensure their heavy investment in funding the institution is performing well.

7. Public Policy Recommendation

Secondly, the gainful employment regulation provides useful learning for regulatory frameworks for virtual schools. Gainful employment regulation is designed to measure the graduating outcomes of for-profit online universities in order to assess whether they add economic value to those receiving their degrees. The most standard assessment of gainful employment is the share of an institution's federal student loan borrowers who default within a 2 or 3 year period after beginning repayment. For-profit online universities have a much higher default rate coming in at 19.1% compared to other types of programs with private not-for-profit institutions coming in at 7.2%, public colleges coming in at 12.9% (Mathew Hodgman, 2018).

There is a wide variety of causation/correlation issues with the regulation that is difficult to avoid in making these comparisons. Firstly, students that choose to enroll in for-profit online universities are disproportionately "second-chance learners" who have normally experienced some kind of failure in traditional education (Nair et al, 2013). Many of the students are working mothers, single mothers and pursuing a degree alongside other responsibilities. Secondly, for-profit online universities are generally unselective in that any student with a high school graduate diploma can participate whereas other types of colleges generally have more stringent entry requirements. Critics would say that these universities take anybody they can to maximize their tuition revenue and profitability and industry proponents would say they provide a valuable last-chance education solution that gives people a flexible way to further their qualifications.

At the core, it is difficult to analyze any counter-factual cases as to what would have happened if a student who attended a for-profit online university attended a different type of institution so most comparisons tend to be larger reflections of differences in cohorts across university. Ideally, one could run randomized control trials to measure the value-add provided by different institutions but this would result in ethical concerns as it would likely compromise education outcomes. One

could also look at natural experiments to compare students who left a for-profit university and attended another type of institution compared to those that stayed although this is again likely to have self-selection issues.

I recommend that regulators focus heavily on two types of measurements to establish a comparable framework to the gainful employment indicator used by universities. Firstly, virtual high schools should be assessed based on the absolute academic achievement of their students. In New Zealand, this would be measured by NCEA pass rates for Level 1, Level 2 and Level 3 and the proportion of students that score achieved, merit, excellence and NZQA Scholarship pass or outstanding. These grades are various standards of achievement. For the relevant country, regulators should consider the equivalent grading framework to assess achievement. While grades are imperfect and do not capture hardships in a student's life or economic adversity, they are comparable between institutions and provide a basis for a reasonable comparable analysis. As discussed earlier, the challenge with absolute grades is that it does not adjust for the calibre of incoming students and therefore penalizes schools that serves relatively underperforming students creating an adverse incentive that drives high performing schools to only serve high performing students subsequently meaning the best teachers are generally serving a nation's relatively best students as opposed to helping weaker students catch up.

As a result, secondly, I recommend a measure of relative achievement. International benchmarked standards such as PISA (Programme for International Student Assessment), SAT or other such standardized assessments are a good example of potential options. Schools should test incoming students and keep a repository of their baseline scores and then measure their value-add or "alpha" to the students. Schools can then develop reputations associated with their ability to serve particular types of students. Teachers that enjoy helping somebody who has fallen off a strong academic path can sort into schools that serve weaker students and teachers that

thrive off helping accelerated students can support these types of learners.

Additional measures of performance the school could consider would be university matriculation rates. Universities vary enormously in quality so ideally a measure of their efficacy in getting students into various types of universities is ideal. Not all students wish to go to a university, such as a four year college in the United States, and for students interested in pursuing vocational pathways and becoming nurses, construction workers, electricians, real estate agents this measure may not adequately reflect a school's adequacy for them.

Another learning from for-profit online universities worth considering for regulators is graduation rate analysis. Many commentators criticize large for-profit online universities for their high drop-out rates. A student may drop out of a university program for a variety of reasons. Firstly, the student may be unable to cope with the work as a result of its difficulty or because of other factors occurring in life that make the work. Secondly, the student may be unable to pay the tuition costs because of unforeseen financial hardship, poor budgeting or other such factors. Thirdly, the student may be unsatisfied with the program. Generally, high dropout rates are likely to be correlated with weaker programs serving lower-income students with weak pastoral support (John Morris, 2019). Public repositories compiled by the government release data on drop-out rates by university and in some cases by program. This helps to warn prospective students of programs that historically have had large churn issues so they can better assess whether or not they will actually finish the qualification they have begun.

In the case of virtual schools, I recommend requiring or encouraging schools to release final graduation rates defined as the proportion of students who start the school that end up completing a high school diploma and also the year over year graduation rates which captures the percentage of students who progress from one year level to the next. If a high proportion of students leave a school

in a specific window it could suggest a weaker program in those years or more compelling outside options relative to the financial cost. These types of statistics are compiled by the US Department of Education for colleges and a similar approach could be applied to high school.

This data being publicly released can create a self-fulfilling prophecy where the best schools with the highest graduation rates become the most attractive and students at schools with low graduation rates feel that this behavior is more normalized and the psychological cost of embarrassment declines. While this is a fair criticism of holding schools to public account, savvy parents can often figure out these numbers with some ranges through careful primary research. Public accountability of schools provides a strong incentive to create high-performing environments from inception and will likely force weaker schools out of operation while stronger schools consolidate their position and serve more students. This natural selection is acceptable with virtual schools because they are substantially more scalable than schools that can only be accessed for a certain geography and as such, winner-take-most dynamics where a dominant online school captures a significant market share is an acceptable outcome. This is particularly true given virtual schools are going to be a consideration for a minority of parents and a plethora of brick-and-mortar schools will continue to be available for the foreseeable future.

In conclusion, for-profit online universities provide a large dataset, substantial public policy innovation and a wide variety of learnings both positive and negative that can guide regulators in informing their public policy outlooks for virtual high schools. While some anomalies such as the principal-agent problem (Hart et al, 1992) and variance in academic success metrics are important to note, in general, this is a highly relevant comparable regulatory example that can help regulators avoid the same pitfall previous states and federal governments have encountered in the United States. The evidence-based findings from this literature could have

useful implications for prospective regulation of the virtual high school industry.

7.6 Transparency of Virtual Schooling Student Outcomes and Teacher Performance

One of the benefits to policymakers that should be a major focus is the enhanced transparency of virtual schools. Typically, a government pays large expenditures to subsidize public school education often with the only visibility of student achievement being student grades and ad-hoc inspection initiatives on an annual basis in which Ministry of Education individuals inspect the school site. Limited visibility into teacher feedback from students or teacher performance on a more ongoing basis exists.

In the virtual schooling environment, with all sessions potentially being recorded, regulators using transcription, sentiment analysis and big-data analysis could screen all lessons occuring in a country for anomalous student-teacher interactions. Examples of this could include if the teacher is talking for a disproportionate amount of a given class or if students are talking for a disproportionate amount of a given class. Sentiment analysis (Ortigosa et al, 2014) could detect clusters of disengaged students based on their facial expressions. Students could give feedback on how they found the usefulness of a lesson after every interaction with a teacher. Governments could build up massive data-sets showing every public teacher in the country's feedback scores from students, teaching habits based on their relative speaking habits, syllable length, fluctuations during the school day and during the school year in style and other such nuances. This would enable governments to far more easily detect underperforming schools and over performing schools and act accordingly. If a data-driven model of high performance teaching can be developed based on regression analysis evaluating ratio of teacher contribution

to class discussion, student satisfaction scores and student academic achievement similar to techniques used in my own cross-sectional and randomized control trial analysis, teachers who deviate from this mould can be targeted for additional professional development with precise feedback on areas of improvement.

Additionally, disciplinary issues are likely to take up a significantly lower share of time because all interactions are recorded, reducing the ability for interactions where the only evidence a teacher and a student has is hearsay. While this may seem trivial, regulators who are responsible for schools in lower-performing environments such as Northland, New Zealand have to spend substantial time on disciplinary issues (Ministry of Education, 2018).

If a regulator such as the New Zealand Ministry of Education was able to form a clear view of school level performance by having a clear picture of the efficacy of the teachers inside the schools, the types of interactions taking place and the engagement levels of students alongside more insight into their academic performance, they are likely to be able to invest behind subsidy programs with more confidence and understanding of the risk factors. By being able to intervene more effectively in underperforming schools and replicating strong initiatives at high performing schools, they are more likely to realize a good social return on investment for their education spending.

With greater transparency comes greater privacy concerns (Lieberman, 2020). If the teaching industry is already experiencing teacher shortages, does recording all teachers, empowering students to provide daily reviews to teachers so they are accountable in every single lesson and forcing teachers to confront a clear view of their student's engagement make the industry more attractive to join or not? Such a structure improves meritocracy and performance tracking which may make the industry relatively more attractive for high performing teachers and relatively less attractive for weaker teachers. Such selection effects may be favorable for average

teaching quality in the industry. Additionally, even if these initiatives reduce teacher recruitment levels, virtual schooling provides much greater access to teachers from a national or global pool and helps to remove the constraints of having high performing teachers available in all locations. A smaller group of high performing teachers having their work be more scalable to students across large areas may be more effective than a larger number of teachers with larger variances in work ethic, performance and skill interacting in opaque delivery environments with students. Many industries such as banks, hedge funds and customer service are regularly recorded for quality control and security purposes. The more vulnerable stakeholder out of the teacher and the child is the child as the teacher opted into this environment and the child is under the age of eighteen so a cost-benefit analysis around the benefits of recording for a government should largely revolve around the child.

Other criticisms of recording of interactions for the purpose of policymakers is that the sheer volume of recordings is massive and the potential storage costs of all of this data non-trivial. Big-data analysis techniques available through vendors like Amazon Web Services and IBM Watson help to deal with this concern. Content could also be deleted after a certain period of time.

Other methods of mitigating privacy concerns would be to keep the transcripts anonymous so a school can receive macro-level insights into how their teachers are performing but cannot identify which particular teacher is responsible for this. Similarly, broad patterns in student engagement in virtual classes could be studied without being able to single out individual contributions through anonymization. The trade-off between quality of insights and actionability of learnings from data and privacy is an important debate that regulators will need to have depending on their political preferences and relevant constituencies.

7.7 Conclusion

In this chapter, I have made two recommendations. Firstly, it is important that policy makers, especially those outside of the United States and the United Kingdom continue to focus on evidence-based public policy design based on robust academic evidence as opposed to political instincts or intuition. While academics in education have made brilliant strides in adopting gold-standard empirical techniques such as randomized control trials in the last thirty years, public policy continues to lag. Secondly, it is important that policy makers proactively release regulatory frameworks to support the evolution of online schooling.

I have made a significant number of recommendations on the regulation of online schools. It will require reasonable resources to monitor these implementations, but fortunately, if regulation is architected correctly, a lot of the necessary data required for compliance can be generated automatically from online high school's virtual classrooms or customer relations management databases directly. If a given regulator does not have sufficient resources to track, monitor and deploy a wide range of regulations on this matter, clear expectations for operators can be laid out with the burden placed on online schools to self-regulate in accordance (or face penalties if they are investigated and found to not be in compliance). Xi Jinping, in his recent regulation of China's for-profit online tutoring segment in July-August 2021, provided a notable example of this approach.

It is obvious that brick-and-mortar schools and online schools cannot be regulated in exactly the same way. For example, many of the requirements of obtaining a school license that require physical buildings which pass various health and safety conditions in a country like New Zealand are not relevant. In general, regulation that focuses on absolute and relative student outcomes, publication of academic achievement data and other outcome-focused measures can be consistent across schooling formats. Many of the practical details around other matters, such as

international student visa requirements, are less relevant in a virtual school where a student may never visit the jurisdiction in which the online school is actually located. Most countries, outside of the USA and UK, have limited to no regulation specifically around online high schools and require any kind of school to follow the same registration requirements which often creates a practical impossibility for online schools to be formed in certain jurisdictions. My suggestions can generally apply to most OECD countries but are particularly applicable in Australia, New Zealand, Canada and the United Kingdom which can quickly benefit from adapting to some of modern regulation and general understanding of the operations of online schools that exist in the United States.

8

Conclusion

In this final chapter, I first summarize the key findings from each chapter to present my argument as a whole. Secondly, I discuss my contributions to academic literature. Thirdly, I consider the limitations of the thesis. Fourthly, I discuss recommendations for future research. Finally, I conclude the thesis with some recommendations on where to from here in general for both education practitioners and the education technology industry.

My key research question was as follows:

What drives student outcomes and student satisfaction within online schools? Specifically, can systematic algorithmic matching between students and teachers using psychometric characteristics and/or gender improve student outcomes and student satisfaction in an online learning environment?

8.1 Primary Findings

I firstly restate some of the key findings from my chapters and offer an executive summary of the public policy recommendations.

In my second chapter, the systematic review of student outcomes and student satisfaction in virtual schools found key drivers to be (1) the impact of parental involvement, (2) the importance of self-motivation, (3) the impact of peer effects, (4) the impact on general student achievement and (5) the importance of mentorship in online learning environments.

In my third chapter, the systematic review of education matching showed (1) a lack of impact of matching by gender and (2) a lack of impact of matching by ethnicity but broadly a sparse literature that needed more contribution. I also noted a deficiency in analysis of personality-based matching.

In my fourth chapter, my qualitative analysis found the following recurring themes in students' perceived requirements of success of virtual learning: (1) self-motivation, (2) desire for academic acceleration through extension coursework, (3) less stressful social and competitive environment (4) flexibility in schedule facilitating other activities. The core themes we found in students' perceived weaknesses of online schooling were: (1) higher barriers for communication with teachers, (2) less spontaneous opportunities for social interaction, (3) limited access to school extracurriculars.

In my fifth chapter, I found that (1) tutor HEXACO traits had a minimal first order impact on average student session ratings, (2) student HEXACO traits agreeableness and openness to experience are positively correlated with increased student ratings and these findings are robust across two independent data-sets, (3) session ratings are positively correlated with HEXACO sub-dimension Sincerity when weighted by square-root of the average number of sessions at the 5% significance level and (4) the relative difference between the HEXACO sub-dimension Inquisitiveness

is positively correlated with average session score.

In my sixth chapter, the randomized control trial produced a number of important findings. The key findings were that (1) matching by the HEXACO algorithm (personality matching) lead to a statistically significant increase in student session ratings in Western Europe and Latin America, (2) matching by the HEXACO algorithm (personality matching) lead to a statistically significant decrease in student session ratings in Asia and Eastern Europe, (3) matching by the HEXACO algorithm lead to a statistically significant improve in SAT writing performance but not on reading or mathematics, (4) gender matching lead to a statistically significant increase in average session rating and (5) gender matching lead to a highly statistically significant impact on the first two lessons of a student-tutor pair and then waned in significance as the number of interactions grew. Other key findings were that (6) student outcomes measured by centred SAT score improvement was not correlated with student satisfaction scores measured by average session rating.

All of my six initial hypotheses were proven wrong by the randomized control trial data. None of the core HEXACO traits proved to have a statistically significant impact when used for absolute difference matching. Additionally, gender matching was shown to have a statistically positive impact on student satisfaction for pairs that had two or fewer interactions.

In my seventh chapter, I made a number of public policy recommendations. I suggested (1) broader adoption of evidence-based outcome measures such as randomized control trials should be used by public policy makers as the status quo, especially in countries like New Zealand. (2) I also suggested public policy makers should proactively develop legislation to support the emerging trend of online high schooling particularly post COVID-19 and gave various suggestions for consideration.

8.2 Contribution to Literature

This thesis makes a notable contribution to the education matching literature. Most of the existing work in this space has involved case study based or small sample size analysis of ethnicity or gender matching. To my knowledge, this randomized control trial is the first of its kind in the education matching literature and found a statistically significant causal impact of matching based on personality in Western Europe and Latin American students on student student satisfaction. It also produced a causal impact on student outcomes as measured by improvement in SAT writing scores. The literature on the use of personality matching was particularly sparse. My findings can encourage more research into the use of matching in a broad array of areas across education between students and teachers but also in doctoral research between students and professors, between lawyers and clients, between coaches and athletes, between romantic couples and other such applications. I have used the HEXACO assessment but a wide variety of other psychometric batteries can be used.

My findings make a notable contribution to gender matching. The gender matching research to date has been ambiguous in determining whether same or different gender matching for either men or women led to any meaningful results. My data would suggest that a clear first impressions bias exists where students report higher satisfaction scores for individuals of the same gender. This insight may help to better explain the mechanism driving the results of other experiments in the field performed to date.

Beyond the specific findings, the experimental design of randomized control trials to evaluate the impact of various algorithmic matching procedures on student outcomes and satisfaction can be replicated across more data-sets, across different types of algorithms, with larger sample sizes. My thesis combined both a cross-sectional analysis and a randomized control trial which enables the evaluation

of statistical models against two separate data-sets and an experimental design that evaluated both correlation and causation relationships. It also provided an example of how to make the most out of a single data-partner with insights from the available database in terms of key variables informing the design of the randomized control trial intervention.

My systematic reviews help to support researchers in more rapidly understanding the state of the online schooling literature in an unbiased manner so more individuals can perform work in this academically sparse but practically highly significant space that has only become more relevant post COVID-19. More broadly, the use of systematic reviews as compared to general literature reviews helps to advance the rigor of the research conducted in the online education field.

My qualitative research focused on the high performing Stanford Online High School provides a useful set of insights that can help researchers better understand the drivers for high performing students to choose virtual schools. It also helps researchers better understand the mechanisms behind the success of high achieving online high schools. Traditionally large scale industry research naturally shines more light on students pursuing credit recovery historically achieving poorly academically as this constitutes the bulk of enrollment in the virtual schooling sector in the United States pre COVID-19.

Finally, my public policy recommendations provide a useful primer for education departments to get up to speed on the key opportunities, risk factors, financing models and relevant case studies in evaluating how they should approach online schooling adoption within their country. When this was first written, virtual schooling was a sector growing with increasing interest but given the need for massive adoption of online schooling on short notice during the COVID-19 pandemic, it would seem that these insights will be sought after more proactively. My conversations with the New Zealand Ministry of Education to date would suggest that governments

are taking this learning model substantially more seriously.

8.3 Limitations of Analysis

As previously stated, my thesis has a variety of weaknesses that can be improved by future contributions to this space. Firstly, my core findings about the use of psychometric matching to enhance student satisfaction (measured by session ratings) and student outcomes (measured by centred SAT score improvement) were obtained on a sample of one-on-one interactions with students and tutors. Virtually all school environments involve a teacher and a large group of students. While my findings are directly applicable to the vast network of online tutoring companies like Chegg, Cluey Learning, VIPKids, Varsity Tutors and other such platforms, more work is required before these findings can be scaled to a full school use-case. One could optimize a set of pair-wise allocations between a given teacher and a set of students and ignore student-to-student interactions as a simplifying assumption when using the algorithm but this would potentially undermine its effectiveness as peer-interactions partially crowd out the student-tutor interactions.

Secondly, the sample size of students with a pre and post assessment on the SAT diagnostics used in the randomized control trial to measure student outcomes, while statistically significant, is only 42 students. The research findings could be strengthened by running the analysis on a larger sample size. Additionally, the sample being used skewed towards more gifted students. In order to test validity to a broader pool of students, a more mixed distribution of academic ability of the students in the sample would be useful.

Thirdly, my personality matching algorithm only considered various coefficients of weighting of relative differences between students and tutors. A more sophisticated algorithm might encompass a mixture of relative differences, absolute differences and absolute scores for students and tutors in assigning matching. It may also

consider the use of logarithms, exponents or higher order pairs. While these additional modifications may be hard to intuitively deduce, they may lead to further impact. A wide variety of other types of matching such as IQ matching and interest matching could be attempted.

Fourthly, my measure of student satisfaction of session scores captured after every lesson are logical but may not be highly correlated with long-term satisfaction scores. My measure of student outcome of SAT score improvement is also not without flaw as no two SAT tests are identical (and using an SAT test twice would create biases as students would remember questions and subsequently outperform). While we correct for this by using tests designed to be of comparable difficulty and subtracting the mean performance on each test, there is still imperfection over a perfectly comparable test. Additionally, while SAT is widely accepted as a major standardized test, it may not capture broader changes in student outcomes such as social skills, confidence levels and development in subjects outside of mathematics and english.

Fifthly, the cohort of analysis for the qualitative review was primarily students from two online high schools (Stanford Online High School and Florida Virtual). These high schools are both considered to be very high performing relative to typical online high schools. Stanford, in particular, is regarded as being very academically rigorous with a ~40% acceptance rate (Tomohiro Hoshi, 2020). Any analysis from these schools is likely to only be relevant to high achieving students engaging in online schools. Students at these schools typically have both high intrinsic motivation, self-control and strong parental engagement. Stanford Online High School students also tend to disproportionately be from higher income families which skews the out-of-sample applicability of the findings. Additionally, while twenty-one is a reasonable sample size for qualitative review a broader based investigation may yield additional insights.

Sixthly, my public policy suggestions around financing models only examined

funding models from the United States education and healthcare system. It did not include alternative financing models like income sharing agreements or FORTE (financing of route to employment). Other large government funded education models like Singapore which provide financing for both domestic and international students should be examined in more detail.

A final limitation is my implicit assumption that schooling is primarily about learning. There are additional considerations such as social skill development, socialization more generally, talent development in areas such as communication skills and time management and also childcare enabling parents to work. This research was conducted through the lens of enhancing student performance but parents are more likely to take a holistic approach. Future research could consider parental engagement and involvement more rigorously or even parental satisfaction levels with education of their children.

8.4 Suggestions for Future Research

There is a wide variety of extension areas for future research. There are a number of enhancements that could be made to the randomized control trial in particular. Firstly, it would be useful to understand why the personality matching algorithm I used had a statistically significant impact on writing but not on reading or mathematics in the SAT. Further randomized control trials could attempt to see whether these findings are replicable or if, with different coefficients, outperformance can be generated in mathematics and reading using different coefficients. Intuitively, it would seem possible that if writing scores can be systematically improved by personality matching with various coefficients of a student and a tutor, different coefficients could yield comparable results across different subjects.

Secondly, it would be useful to build regionally optimized matching algorithms that consistently generate positive increases in student outcomes and student

sessions for specific ethnicities or cultural backgrounds which my research suggests produces meaningful variance.

Thirdly, it would be useful to run a larger n-value analysis of student outcome improvement as my sample size was meaningful but could benefit from being larger.

Fourthly, other types of personality psychometric assessments could be considered beyond HEXACO such as Big-5, the preceding test to HEXACO.

Fifthly, other measures of student satisfaction and student outcomes could be used. The use of average session rating is a good proxy for satisfaction during a given session but do we know that a string of satisfied sessions necessitates a satisfied student at the end of the term? Satisfaction is likely to be an endogenous variable affected by perceived and realized student outcomes over different time horizons. Other tests beyond the SAT could be used to measure student outcomes such as the IELTS, TOEFL, GMAT, GRE, PAT or other subject specific curriculum such as A Levels, SAT Subject Tests or the International Baccalaureate.

Sixthly, other age groups could be assessed including younger students (in kindergarten, primary or intermediate school) or adult students pursuing graduate education or second-chance learners who are completing academic milestones later than most individuals their age.

Finally, with regards to the randomized control trial, I use region as a proxy for cultural differences but a dataset that accurately tags each individual by ethnicity would yield more clear findings because, for example, analysis could be conducted for individuals who declare themselves as Chinese as opposed to using a proxy of China (which has a diverse mix of cultural backgrounds with a large proportion of expatriates).

I recommend further empirical researchers in the education space adopt the combined use of the cross-sectional analysis on historical data followed by a randomized control trial as it provides a compelling opportunity to test initial

findings on a fresh sample while also providing causal findings as opposed to purely correlation analysis.

COVID-19 presents a huge natural experiment that would make economists giddy with the extent of research possibilities. A broad survey of students' experiences learning online during COVID would now be able to be conducted at a substantially lower cost because it would be orders of magnitude easier to find relevant samples. Additionally, while the students I researched had already opted into online schooling of their own accord and subsequently are likely to be skewed by a form of survivorship-bias (for staying in this schooling model), many learners during COVID-19 had to participate in online learning with no choice. As a result, they may offer a more neutral perspective into the potential benefits and costs of virtual schooling.

A survey of principals and teachers around their experiences with online education would also be beneficial. Humans tend to be inherently scared of change and asking a teacher about their views of online education prior to trying isn't likely to yield anything usable. Now, many teachers have been forced to try teaching online and the insights available from this experience have profound significance for how easy it will be to recruit teachers to virtual schools in the future. If the experience was strong, virtual schools may flourish with a teaching force who are skeptics born again. If the experience was negative, virtual schools may fight an uphill battle in recruiting quality teaching staff.

The parent voice must also be included. Many parents during COVID-19 were forced to get closer to their child's education than they ever have before as they lived within the same roof during schooling hours for more than a month in many OECD countries. Asking parents about their experiences with online education, their perception of their child's engagement levels and the impact online schooling had on the family will offer significant insights. This is of crucial importance given the conclusive finding from my systematic review that parental engagement is a

major driver of student outcomes in virtual schools.

Beyond surveys, studying the actual shifts in the education landscape will be illuminating. Will leading schools like Phillips Andover and Phillips Exeter continue to offer an online learning component to their programs? Will there be a spike in adoption of online schooling both part-time and full-time? Will online tutoring companies take share from brick-and-mortar tutoring companies? Will more students transfer to online university degrees over high cost in-person experiences? Answering these questions will yield profound insights into the future of space.

8.5 Conclusion

The arguments and evidence presented in this thesis are relevant to a number of important developments in education across both the public and private sector. The first major consideration is the need for rigorous evidence-based assessment in education through gold-standard experimental designs such as randomized control trials. The second major consideration raised by the thesis is the need to consider proactive online schooling legislation to prepare for the growth in this type of schooling, accelerated by COVID-19.

Throughout my thesis, I have used a mixed-methods approach ranging from qualitative analysis when it came to assessing drivers for students to attend online high school to cross-sectional analysis and randomized control trial techniques when the data-set and population in question allowed for such approaches and assessed the existing body of evidence through a systematic review. Some aspects of education cannot easily be assessed through gold-standard techniques like randomized controlled trials because of practical restrictions like the age of participants or the potential negative consequences of the intervention. I advocate for a more careful consideration of the most robust experimental design technique possible for a given question, with a bias towards quantitative studies.

Future research should examine the novel psychometric matching findings uncovered. By considering different psychometric matching algorithms and potentially calibrating them by culture or geographic region of learner, student outcomes and student satisfaction may be meaningfully enhanced for limited marginal cost.

COVID-19 has thrust online schooling into the public domain but it had already been growing as a serious consideration for parents in America for two decades. A virtual schooling divide exists in which countries outside of the United States have very few or no operators meaning student choice is limited. Regulators who are quick to create robust legislation that is fit for the constraints and challenges of

virtual schooling with an awareness of the key drivers of student satisfaction and outcomes will be able to offer students and parents a powerful pathway to access more learning opportunities. The future is bright for countries that commit to a blended learning pathway for students with flexibility across physical campuses to online courses to combinations with appropriate regulation and constraints.

Appendices

A

Chapter Four Appendix

Profiles of People Interviewed

I provide executive summaries of the profiles of the students interviewed to provide further context as to the research sample:

1. A rising 17 year old senior who had spent seven years in brick-and-mortar schools and six years in Stanford Online High school on a full time basis. The individual was highly self-motivated and curious.
2. A 17 year old who had spent 10 years homeschooled and three years in Stanford Online High School. They had never attended a physical school. The student was an elite fencing competitor at national level in America.
3. A 16 year old rising junior who spent 11 years in brick-and-mortar school in the Philippines with two of those years doing Stanford Online High School in

a part-time capacity before going full-time. Previously the student had spent one year in a different online high school part-time.

4. A 17 year old rising senior who was home-schooled from kindergarten to eighth grade. Been attending Stanford Online High School for three years. The student is seeking to pursue a career as an actress and pursues acting while in school.
5. A 17 year old rising senior who has been full-time in Stanford Online High School for three years who was previously an intense ballet dancer.
6. A 17 year old rising senior who had spent four years in Stanford Online High School, five years in a home-schooling environment and 9 years in a brick-and-mortar school.
7. A 17 year old rising senior who has spent six years in Stanford Online High School, one year taking a single course and five years doing full online high school.
8. A 17 year old rising senior who attended public school until fourth grade, was home schooled for fifth grade, blended learning (combining home school and Stanford Online High School) from sixth grade to eight grade and full-time Stanford Online High School from ninth grade to the end of high school.
9. A 14 year old rising sophomore who attended brick-and-mortar school until nineth grade and then moved to full-time Stanford online high school but continues to take one or two offline classes to maintain social interaction. The student has only spent one full year in Stanford Online High school but plans to graduate from it.
10. A 15 year old rising sophomore who switched from homeschooling to private schools on and off regularly moving around the world. When homeschooled, supplementary classes were consumed online. Three years so far at Stanford Online High School with a plan to graduate from it.
11. A 14 year old rising sophomore who spent first and second years of school at

a brick-and-mortar school before being home schooled for the last nine years. She has spent one year at Stanford Online High School and plans to graduate from Stanford Online High School.

12. A 14 year old rising sophomore who was in a local Taiwanese school for three years, two years in home school, two more years in brick-and-mortar and has done seventh grade onwards at Stanford Online High School. She has spent four years so far at Stanford Online High School.
13. A 17 year old senior at Colégio Marista Arquidiocesano. The student did brick-and-mortar in Florida, went to Madrid, Spain, and then moved to Connecticut. The student has been attending school in Sao Paulo, Brazil since 2009. The student is using Stanford Online High school part-time to supplement their traditional school.
14. A 15 year old sophomore at Florida Virtual High School who was homeschooled for the first four years of primary school and was then homeschooled until freshman year when she switched to full-time online school.
15. A 16 year old rising junior at Florida Virtual High School who has attended brick-and-mortar high schools and primary schools but has been taking part-time supplementary classes for the last two years.
16. A 17 year old rising senior at Westlake Boys High School who has been in brick-and-mortar public school in New Zealand but has taken part-time online high school for the last three years.
17. A 17 year old rising senior at Stanford Online High School who attended a physical private high school for eight years before going full-time online school. She is a competitive tennis player.
18. A 13 year old student who has attended private boarding schools for the last 3 years in the USA and five years of private school in China. The student is now evaluating switching to online high schools with their family.
19. A 15 year old student who lives in a remote area of New Zealand and has

attended brick-and-mortar high school so far but is evaluating a switch to online high school part-time or full-time to accelerate her learning.

20. A 16 year old rising junior who has attended a public brick-and-mortar school in New Zealand for all their life but has pursued part-time online instruction for the last two years.
21. A 14 year old student who has attended brick-and-mortar school in Germany and China and has spent the last year full-time in Stanford Online High School.

B

Chapter Five Appendix

Table B.1: Number of Cancelled Sessions

	Cancelled Sessions	Sessions Cancelled by Tutor
Not Cancelled	70691	73569
Cancelled	4752	1874

Table B.2: Distribution of Session Rating (Excluding Old Rating System)

Lower Quartile	4.00
Mean	4.66
Median	5.00
Upper Quartile	5.00
Number of Valid Ratings	7308.00
Number of Unrated Sessions	68135.00

Table B.3: Number of Pairs by Student and Tutor Gender

Gender	Number of Pairs	
	Student Gender	Tutor Gender
N/A	9688	9864
Female	1221	1197
Gender diverse/non-binary	15	4
Male	1262	1132
Other	11	0

Table B.4: Number of Pairs by Student and Tutor Country

Student Country	Number of Pairs	Tutor Country	Number of Pairs
#N/A	1216	N/A	4711
Australia	2257	Australia	1503
New Zealand	4527	New Zealand	2719
Other	2508	Other	771
Singapore	647	Singapore	450
Thailand	467	United Kingdom	387
United States	575	United States	1656

Table B.5: Descriptive Statistics on Unique Pairs

		Percentile			Number of Pairs	
	Mean	25th	50th	75th	Valid Values	N/A Values
Old + New Rating System						
Avg Rating	4.67	4.4	5.0	5	7104	5093
Number of Ratings	6.00	2.0	4.0	8	7104	5093
Number of Sessions	7.43	2.0	5.0	10	11263	934
Number of Cancellations	0.47	0.0	0.0	1	11263	934
Number of Tutor Cancellations	0.21	0.0	0.0	0	11263	934
New Rating System Only						
Avg Rating	4.60	4.0	4.9	5	2457	9740
Number of Ratings	5.83	2.0	4.0	8	2457	9740
Number of Sessions	5.04	1.0	3.0	6	10370	1827

Number of Cancellations	0.51	0.0	0.0	1	10370	1827
Number of Tutor Cancellations	0.22	0.0	0.0	0	10370	1827

Table B.6: Distribution of Age Across Sessions

	Student Age	Tutor Age	(Student - Tutor) Age
Lower Quartile	15.77	19.46	-10.72
Median	16.66	21.69	-6.34
Mean	16.18	22.85	-7.46
Upper Quartile	17.50	25.42	-3.08
# N/A	6660.00	8118.00	10278.00

Table B.7: Sessions by Rating Scheme

Rating Scheme	Freq
	68135
NEW	7308
OLD	33307
NA	39

Table B.8: Sessions by Completion Status and Rating System

	Rating System	
	Old + New	New Only
cancelled	4766	4752
completed	69779	36577
confirmed	677	589
declined	3804	3804
deleted	16445	16445
rescheduled	1630	1630
tentative	11649	11646
NA	39	0

Table 18 shows the minimum, maximum, mean, median, 1st and 3rd quartile across all of the HEXACO's core dimensions, and sub-dimensions for students and tutors.

Table B.9: Distribution of Student and Tutor HEXACO Across Unique Pairs

	Student Percentile				Tutor Percentile			
Variable	Mean	25th	50th	75th	Mean	25th	50th	75th
Aesthetic Appreciation	3.33	2.75	3.50	4.00	3.67	3.25	3.75	4.25
Agreeableness	3.06	2.63	3.13	3.44	3.21	2.81	3.19	3.63
Altruism	3.92	3.50	4.00	4.50	4.02	3.75	4.00	4.50
Anxiety	3.64	3.25	3.75	4.25	3.41	2.75	3.50	4.00
Conscientiousness	3.66	3.25	3.69	4.06	3.86	3.50	3.88	4.25
Creativity	3.56	3.00	3.75	4.25	3.71	3.00	3.75	4.50
Dependence	3.01	2.50	3.00	3.75	3.16	2.50	3.25	3.75
Diligence	4.02	3.50	4.00	4.50	4.27	4.00	4.25	4.75
Emotionality	3.26	2.88	3.25	3.63	3.22	2.81	3.25	3.63
eXtraversion	3.47	3.06	3.56	3.94	3.69	3.31	3.75	4.13
Fairness	3.80	3.25	4.00	4.50	3.89	3.50	4.00	4.50
Fearfulness	2.94	2.25	3.00	3.50	2.85	2.25	3.00	3.25
Flexibility	2.93	2.50	3.00	3.50	3.18	2.75	3.25	3.75
Forgivingness	2.82	2.25	2.75	3.50	2.90	2.25	2.75	3.50
Gentleness	3.24	2.75	3.25	3.75	3.22	2.75	3.50	3.75
Greed Avoidance	2.87	2.25	3.00	3.50	3.18	2.50	3.25	3.75
Honesty Humility	3.35	3.00	3.44	3.81	3.48	3.13	3.44	3.81
Inattentive	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
Inquisitiveness	3.52	3.00	3.50	4.25	3.91	3.50	4.00	4.50
Liveliness	3.47	3.00	3.50	4.00	3.63	3.25	3.75	4.25
Modesty	3.45	3.00	3.50	4.00	3.54	3.00	3.50	4.00
Openness to Experience	3.52	3.13	3.56	3.94	3.77	3.44	3.81	4.13
Organization	3.46	2.75	3.50	4.00	3.75	3.25	3.75	4.50
Patience	3.25	2.50	3.25	4.00	3.53	3.00	3.50	4.00
Perfectionism	3.73	3.25	3.75	4.25	3.81	3.50	4.00	4.25
Prudence	3.42	3.00	3.50	4.00	3.60	3.25	3.75	4.00
Sentimentality	3.43	3.00	3.50	4.00	3.46	3.00	3.50	4.00
Sincerity	3.26	2.75	3.25	4.00	3.31	2.75	3.25	4.00
Sociability	3.55	3.00	3.75	4.00	3.72	3.25	3.75	4.25
Social Boldness	3.27	2.75	3.25	3.75	3.48	3.00	3.50	4.00
Social Self Esteem	3.57	3.00	3.75	4.25	3.91	3.50	4.00	4.50
Unconventionality	3.65	3.25	3.75	4.00	3.79	3.50	3.75	4.25

Number of pairs with valid student HEXACO: 4623

Number of pairs with no student HEXACO: 7574
Number of pairs with valid tutor HEXACO: 6805
Number of pairs with no tutor HEXACO: 5392

Table 19 shows the student activation date when students joined the Crimson Education tutoring platform with the average student joining on 26th March 2018 in this data-set, suggesting an average tenure on the platform of approximately ~20 months.

Table B.10: Distribution of Student Activation Date Across Unique Pairs

Lower Quartile	Mean	Median	Upper Quartile
2017-08-08	2018-03-04	2018-03-06	2018-09-03

Table B.11: |Student - Tutor| HEXACO Factors and Average Session Rating

	Dependent variable:		
	Cross Sectional Average Rating b (SE)		RCT Average Rating b (SE)
	(1)	(2)	(3)
Agreeableness	−0.058*	−0.042	−0.0005
	(0.031)	(0.029)	(0.023)
Constant	4.644***	4.681***	4.666***
	(0.025)	(0.023)	(0.018)
Observations	1,036	1,036	1,970
R^2	0.003	0.002	0.00000
Adjusted R^2	0.002	0.001	−0.001
F Statistic	3.401*	2.063	0.0004

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model

Table B.12: |Student - Tutor| HEXACO Facets and Average Session Rating

	Dependent variable:
	RCT Average Rating b (SE)
Aesthetic Appreciation	−0.006 (0.044)
Altruism	−0.044 (0.058)
Anxiety	0.023 (0.043)
Creativity	0.013 (0.046)
Dependence	−0.022 (0.041)
Diligence	0.040 (0.056)
Fairness	−0.038 (0.046)
Fearfulness	0.009 (0.046)
Flexibility	0.006 (0.048)
Forgivingness	0.008 (0.043)
Gentleness	0.025 (0.052)
Greed Avoidance	0.003 (0.039)
Inquisitiveness	−0.030 (0.046)
Liveliness	−0.006 (0.046)
Modesty	0.008 (0.051)
Organization	−0.036 (0.044)
Patience	−0.019 (0.044)
Perfectionism	0.039 (0.048)
Prudence	−0.010 (0.054)
Sentimentality	−0.001 (0.044)
Sincerity	0.016 (0.047)
Sociability	−0.024 (0.045)
Social Boldness	−0.051 (0.044)
Social Self Esteem	0.015 (0.048)
Unconventionality	−0.021 (0.057)
Constant	4.677*** (0.140)
Observations	338
R^2	0.023
Adjusted R^2	−0.055
F Statistic	0.294 (df = 25; 312)
Note:	*p<0.1; **p<0.05; ***p<0.01 Effect size is unstandardised beta.

Table B.13: Cancellation Rate and Average Session Rating

	Dependent variable:			
	RCT Average Rating			
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
Cancellation Rate	0.101	0.157*		
	(0.094)	(0.094)		
Tutor Cancellation Rate			0.085	0.133
			(0.121)	(0.123)
Constant	4.596***	4.631***	4.599***	4.635***
	(0.011)	(0.011)	(0.011)	(0.010)
Observations	2,457	2,457	2,457	2,457
R^2	0.0005	0.001	0.0002	0.0005
Adjusted R^2	0.0001	0.001	−0.0002	0.0001
F Statistic	1.153	2.793*	0.495	1.176

Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

(1) Unweighted Model (2) Weighted Model (3) Unweighted Model (4) Weighted Model

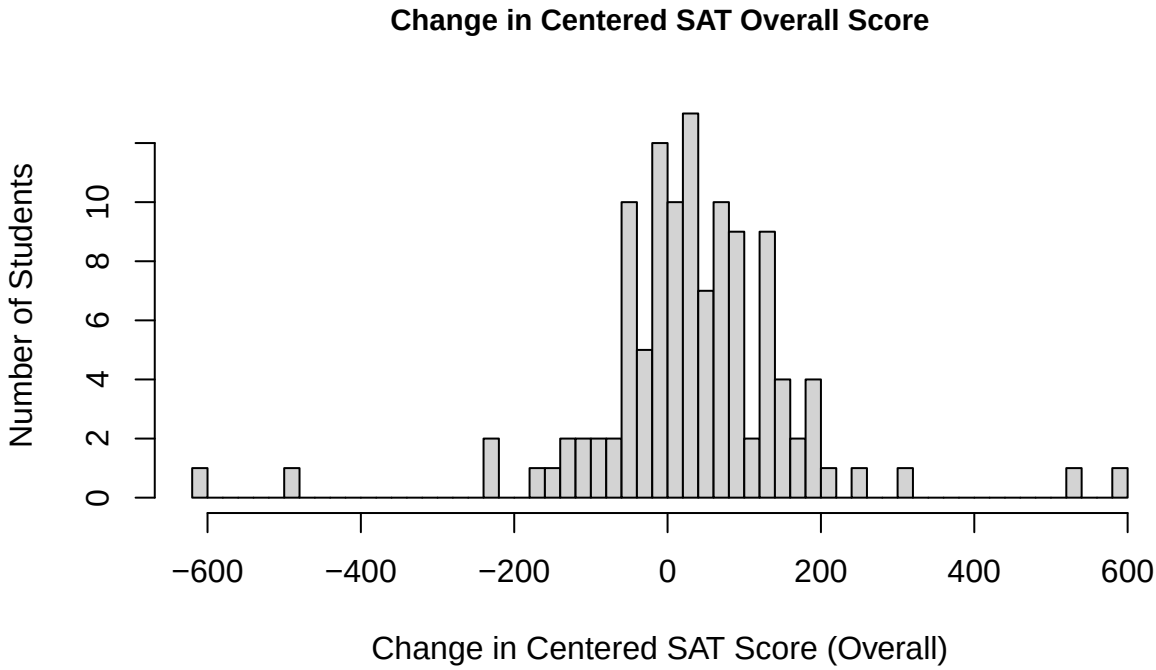
Table B.14: Student HEXACO Facets and Cancellation Rate

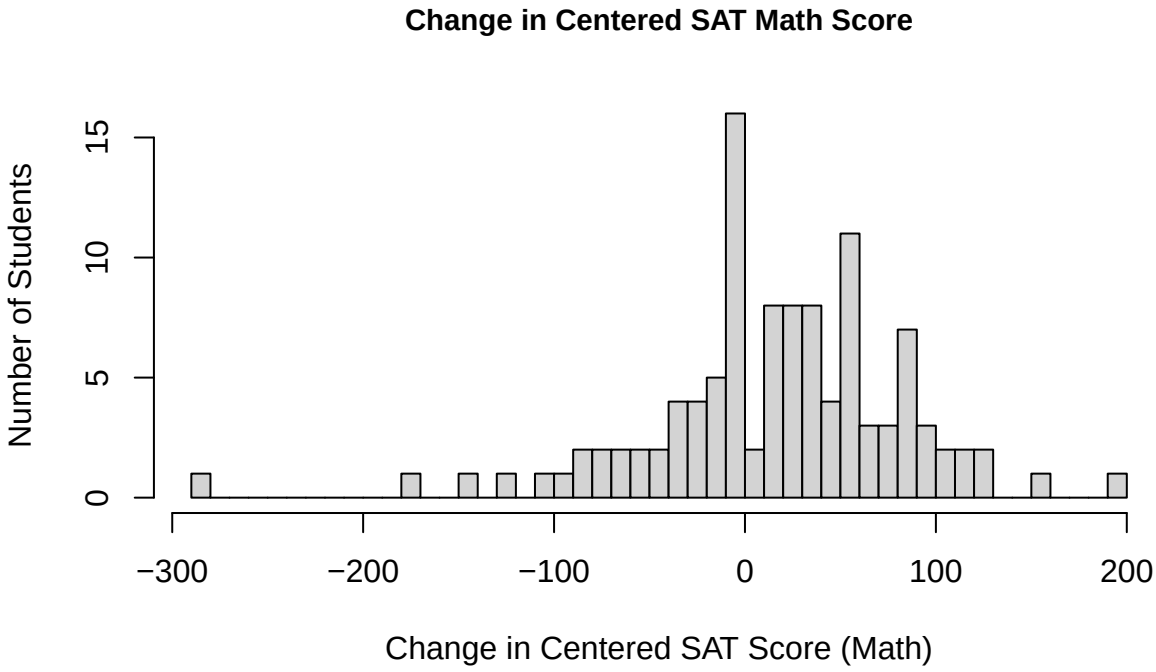
	Dependent variable:	
	Cancellation Rate b (SE)	Tutor Cancellation Rate b (SE)
Dependence	-0.097^{***}	
	(0.032)	
Flexibility	0.083^{**}	
	(0.038)	
Inattentive	1.739^{***}	
	(0.519)	
Inquisitiveness	-0.091^{***}	
	(0.032)	
Liveliness	0.088^{***}	
	(0.033)	
Modesty	0.074^{**}	
	(0.033)	
Organization	0.204^{***}	
	(0.033)	
Prudence	-0.171^{***}	
	(0.040)	
Greed Avoidance		-0.071^{*}
		(0.039)
Constant	-2.914^{***}	-3.329^{***}
	(0.249)	(0.116)
Observations	23,883	28,207
Log Likelihood	−6,007.156	−3,640.583
Akaike Inf. Crit.	12,032.310	7,285.166

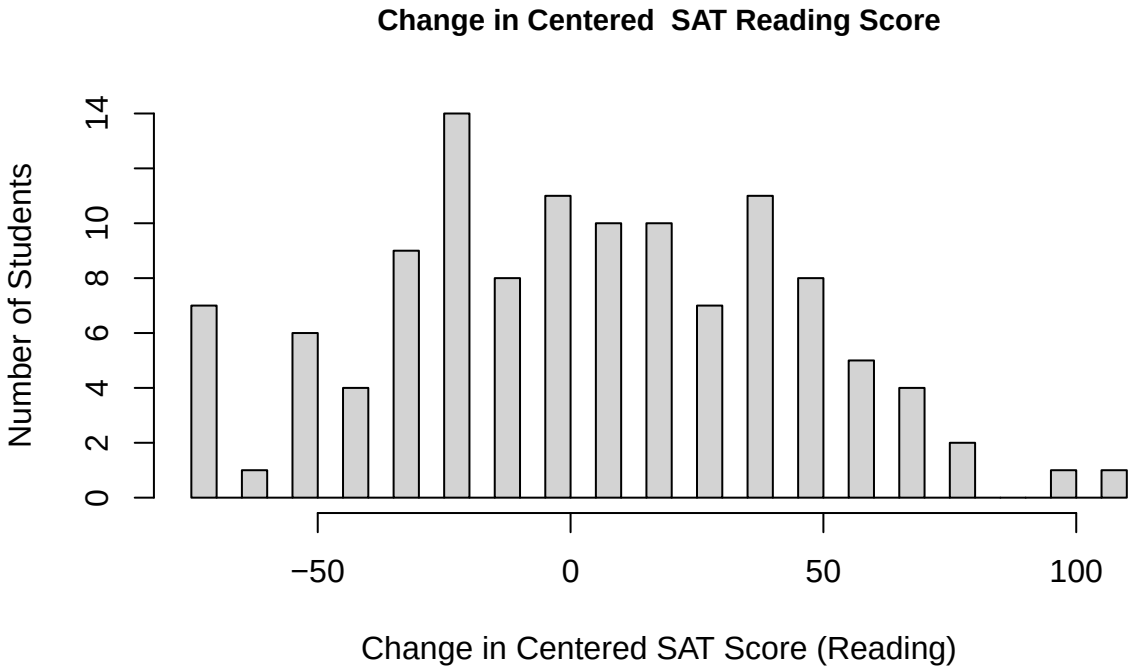
Note: *p<0.1; **p<0.05; ***p<0.01
Effect size is unstandardised beta.

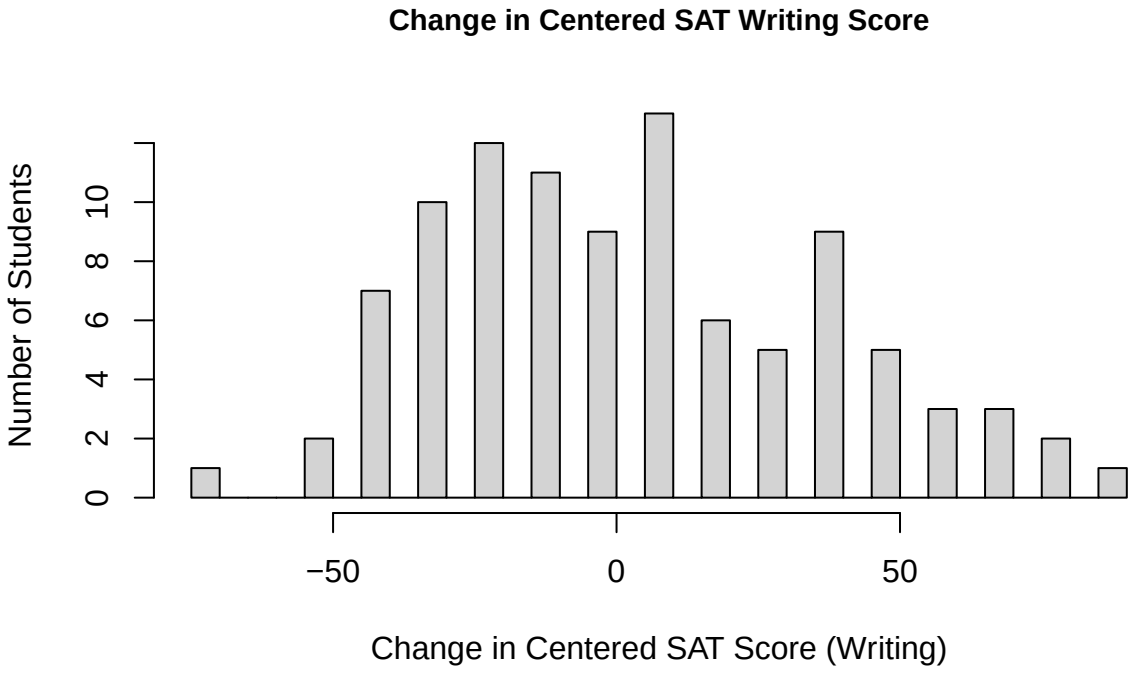
C

Chapter Six Appendix


Change in Centered SAT Overall Score
Number of Students
0
2
4
6
8
10
−600
−400
−200
0
200
400
600
Change in Centered SAT Score (Overall)


Change in Centered SAT Math Score
Number of Students
0
5
10
15
−300
−200
−100
0
100
200
Change in Centered SAT Score (Math)


Change in Centered SAT Reading Score
Number of Students
0
2
4
6
8
10
14
−50
0
50
100
Change in Centered SAT Score (Reading)


Change in Centered SAT Writing Score
Number of Students
0
2
4
6
8
10
−50
0
50
Change in Centered SAT Score (Writing)

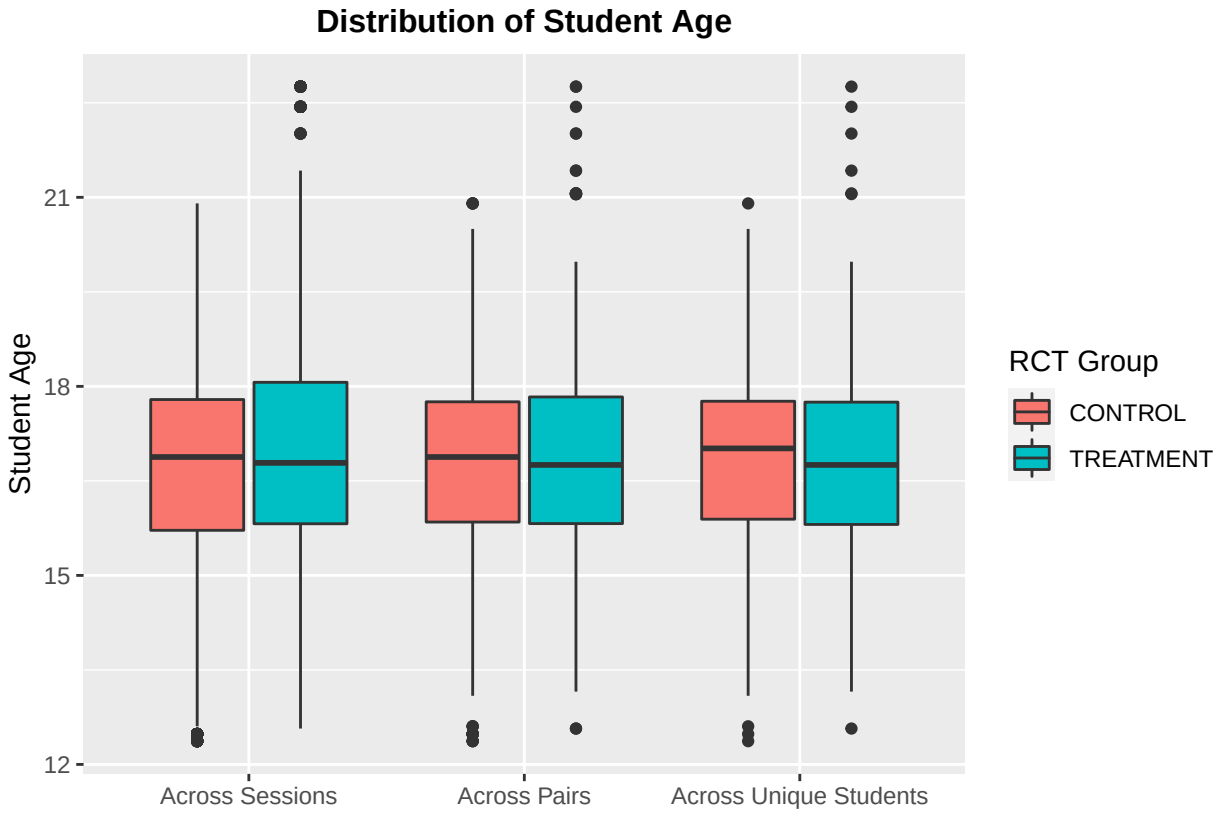


Table C.1: Descriptive Statistics on Session Ratings (Across Unique Pairs, Control vs Treatment)

			Number of Pairs	
	Average Rating	Number of Ratings	With Ratings	Without Ratings
Control	4.669	5013	1008	1050
Treatment	4.663	4729	964	1076

Table C.2: Number of Sessions, Pairs and Students by Student Region Across The Control and Intervention Group

Student Region	Sessions		Pairs		Unique Students	
	Control	Treatment	Control	Treatment	Control	Treatment
Australia	2559	1929	362	259	102	83
East / South-East / South Asia	1936	2112	274	281	72	81
Eastern Europe & Central Asia	2850	2329	278	247	58	48
Latin America	540	789	94	129	31	34
Middle East & Africa	891	1136	119	149	31	32
New Zealand	1210	1778	209	277	54	73
North America	1300	1453	186	237	51	61
Singapore	1026	940	178	168	50	49
Western Europe	1558	962	251	189	67	61
NA	695	602	107	104	27	30

Table C.3: Pearson's Chi-Squared Tests of Independence on Student Region (Control vs Treatment Group)

	Test Statistic	P Value	df
Individual Sessions	482.8	2.778e-98	9
Unique Pairs	52.51	3.624e-08	9
Unique Students	7.69	0.5657	9

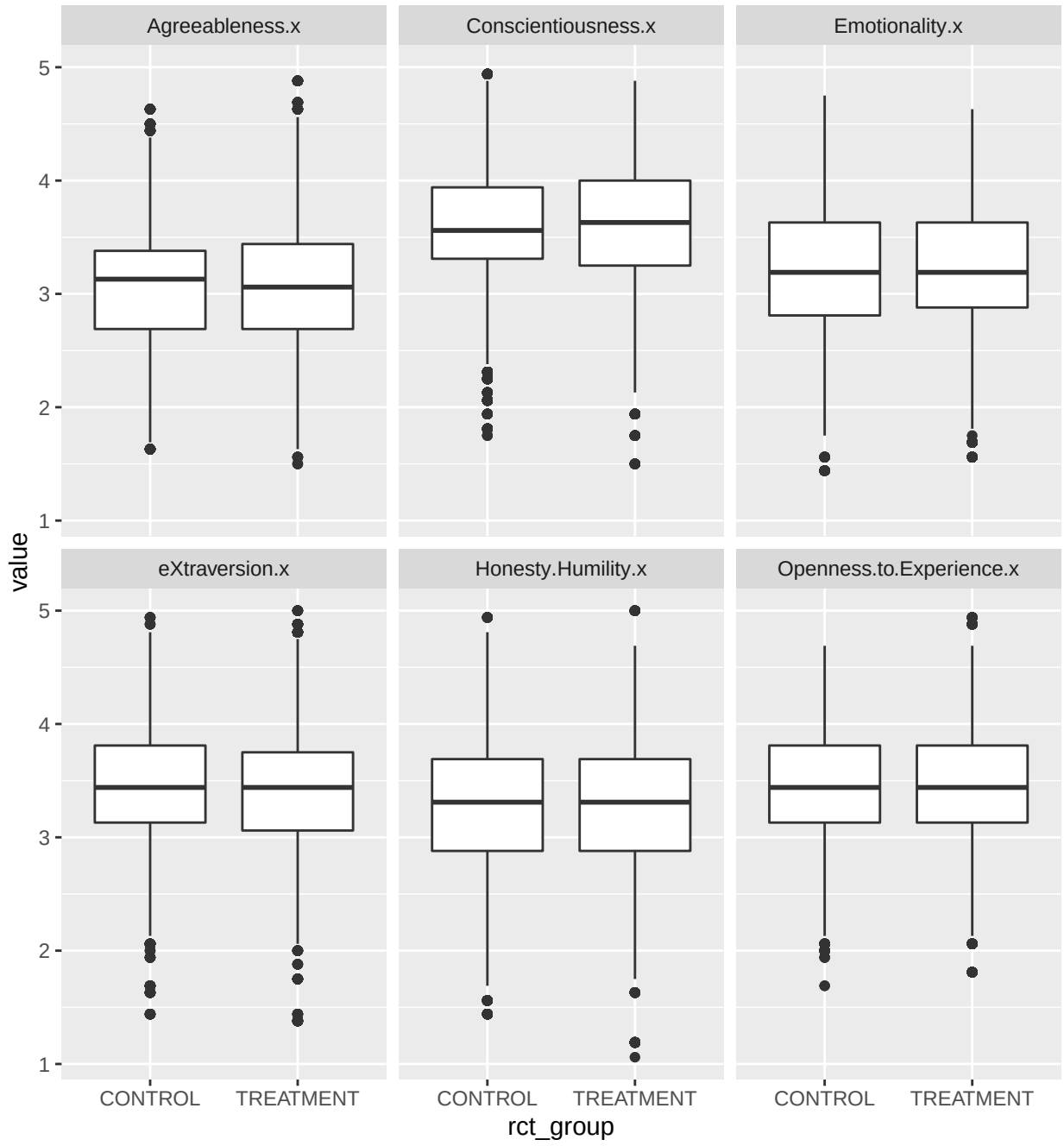
Table C.4: Summary Table of Student HEXACO (Across Individual Sessions)

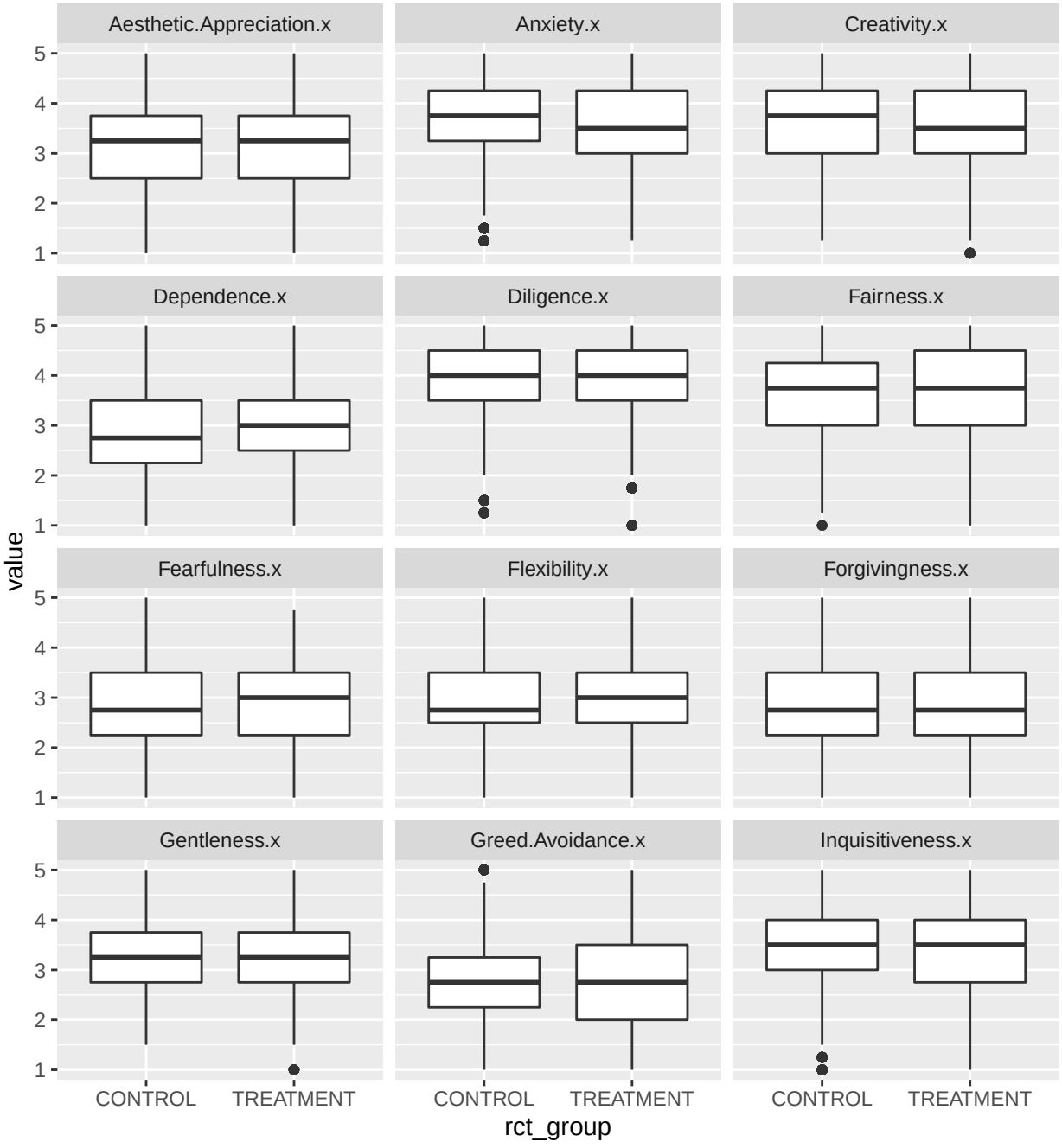
		Individual Sessions		Unique Pairs		Unique Students	
Variable	RCT Group	Mean	S.Dev	Mean	S.Dev	Mean	S.Dev
Aesthetic Appreciation	control	3.22	0.84	3.22	0.85	3.25	0.86
Aesthetic Appreciation	treatment	3.21	0.83	3.17	0.83	3.18	0.85
Agreeableness	control	3.05	0.54	3.06	0.55	3.07	0.56
Agreeableness	treatment	3.05	0.59	3.08	0.60	3.08	0.60
Anxiety	control	3.69	0.82	3.62	0.83	3.67	0.81

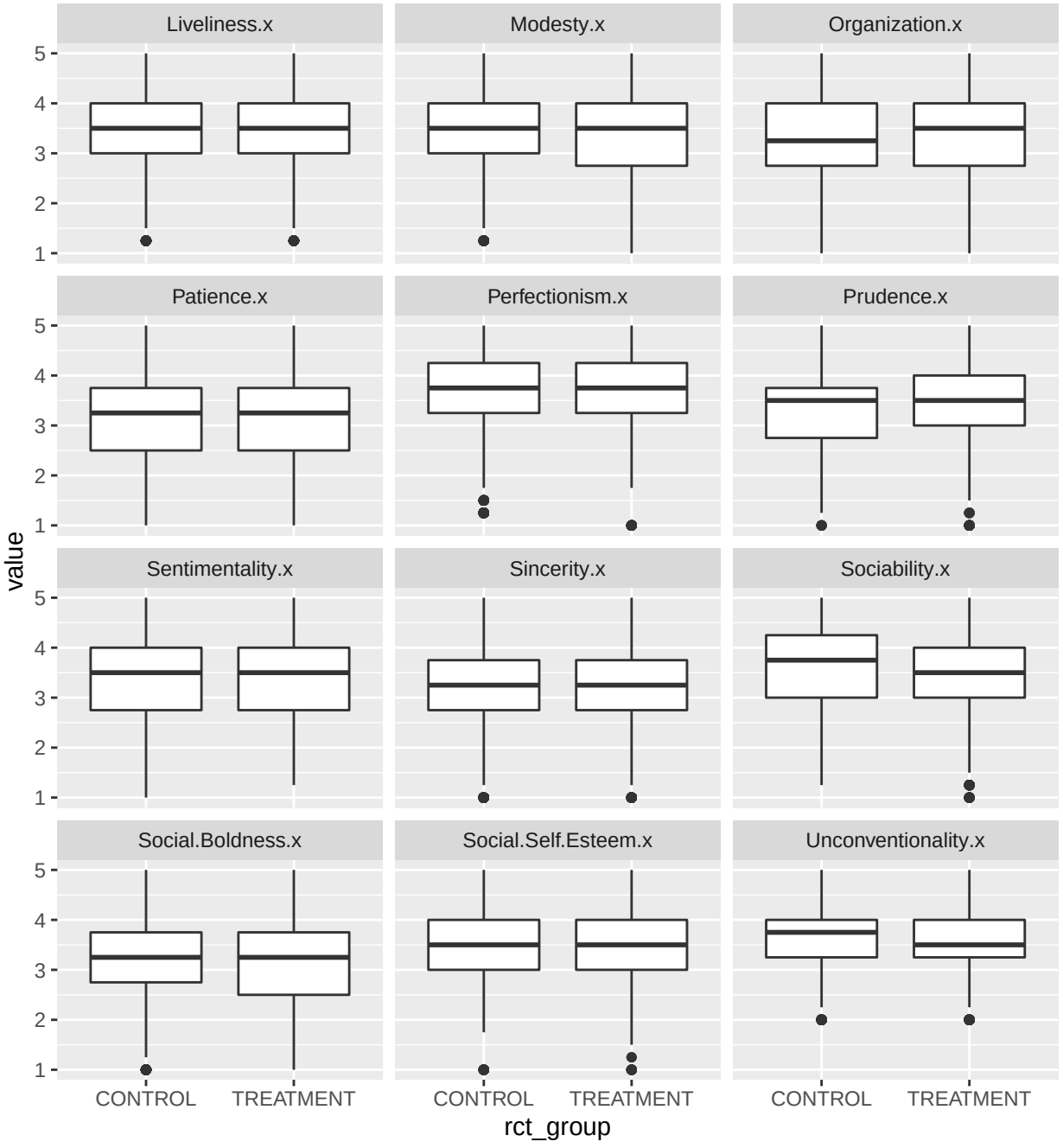
Anxiety	treatment	3.58	0.80	3.58	0.80	3.55	0.79
Conscientiousness	control	3.56	0.57	3.58	0.58	3.60	0.58
Conscientiousness	treatment	3.60	0.56	3.60	0.57	3.61	0.57
Creativity	control	3.59	0.84	3.59	0.84	3.57	0.86
Creativity	treatment	3.59	0.82	3.54	0.84	3.53	0.85
Dependence	control	2.93	0.87	2.95	0.88	3.00	0.87
Dependence	treatment	3.04	0.83	3.01	0.86	3.02	0.88
Diligence	control	3.97	0.73	4.01	0.71	4.01	0.68
Diligence	treatment	3.95	0.69	3.94	0.71	3.96	0.71
Emotionality	control	3.22	0.57	3.20	0.59	3.23	0.58
Emotionality	treatment	3.22	0.54	3.21	0.54	3.21	0.55
eXtraversion	control	3.43	0.57	3.43	0.59	3.45	0.59
eXtraversion	treatment	3.43	0.58	3.45	0.61	3.45	0.60
Fairness	control	3.68	0.83	3.70	0.86	3.75	0.86
Fairness	treatment	3.71	0.90	3.70	0.91	3.73	0.90
Fearfulness	control	2.86	0.79	2.86	0.81	2.84	0.80
Fearfulness	treatment	2.88	0.76	2.87	0.76	2.87	0.76
Flexibility	control	2.90	0.71	2.93	0.72	2.94	0.73
Flexibility	treatment	2.94	0.76	2.98	0.79	3.00	0.79
Forgivingness	control	2.83	0.76	2.83	0.77	2.84	0.78
Forgivingness	treatment	2.83	0.78	2.82	0.80	2.79	0.79
Gentleness	control	3.29	0.67	3.26	0.68	3.27	0.70
Gentleness	treatment	3.25	0.71	3.29	0.70	3.29	0.71
Greed Avoidance	control	2.75	0.91	2.79	0.93	2.82	0.94
Greed Avoidance	treatment	2.77	0.88	2.78	0.89	2.79	0.87
Honesty Humility	control	3.27	0.59	3.29	0.61	3.33	0.61
Honesty Humility	treatment	3.29	0.58	3.29	0.58	3.31	0.58
Inquisitiveness	control	3.40	0.77	3.38	0.80	3.37	0.81
Inquisitiveness	treatment	3.38	0.81	3.37	0.83	3.38	0.83
Liveliness	control	3.45	0.76	3.45	0.77	3.47	0.79
Liveliness	treatment	3.48	0.75	3.50	0.79	3.48	0.80
Modesty	control	3.46	0.74	3.48	0.76	3.51	0.78
Modesty	treatment	3.38	0.76	3.43	0.77	3.47	0.77
Openness to Experience	control	3.46	0.52	3.45	0.53	3.45	0.54
Openness to Experience	treatment	3.45	0.53	3.43	0.54	3.43	0.54
Organization	control	3.32	0.85	3.33	0.89	3.34	0.91
Organization	treatment	3.39	0.84	3.39	0.86	3.39	0.86
Patience	control	3.18	0.81	3.21	0.84	3.23	0.85
Patience	treatment	3.19	0.88	3.23	0.89	3.23	0.88
Perfectionism	control	3.65	0.75	3.65	0.74	3.67	0.74

Perfectionism	treatment	3.66	0.75	3.66	0.75	3.67	0.75
Prudence	control	3.31	0.74	3.34	0.78	3.36	0.76
Prudence	treatment	3.40	0.70	3.40	0.71	3.40	0.72
Sentimentality	control	3.38	0.79	3.36	0.81	3.39	0.81
Sentimentality	treatment	3.38	0.77	3.36	0.79	3.38	0.79
Sincerity	control	3.19	0.82	3.20	0.82	3.23	0.81
Sincerity	treatment	3.28	0.75	3.26	0.78	3.24	0.78
Sociability	control	3.56	0.77	3.52	0.77	3.56	0.77
Sociability	treatment	3.54	0.77	3.55	0.81	3.56	0.80
Social Boldness	control	3.21	0.77	3.20	0.80	3.20	0.81
Social Boldness	treatment	3.22	0.83	3.23	0.83	3.25	0.82
Social Self Esteem	control	3.51	0.68	3.54	0.70	3.55	0.71
Social Self Esteem	treatment	3.47	0.69	3.50	0.69	3.51	0.71
Unconventionality	control	3.63	0.57	3.63	0.56	3.61	0.58
Unconventionality	treatment	3.62	0.59	3.63	0.60	3.64	0.61

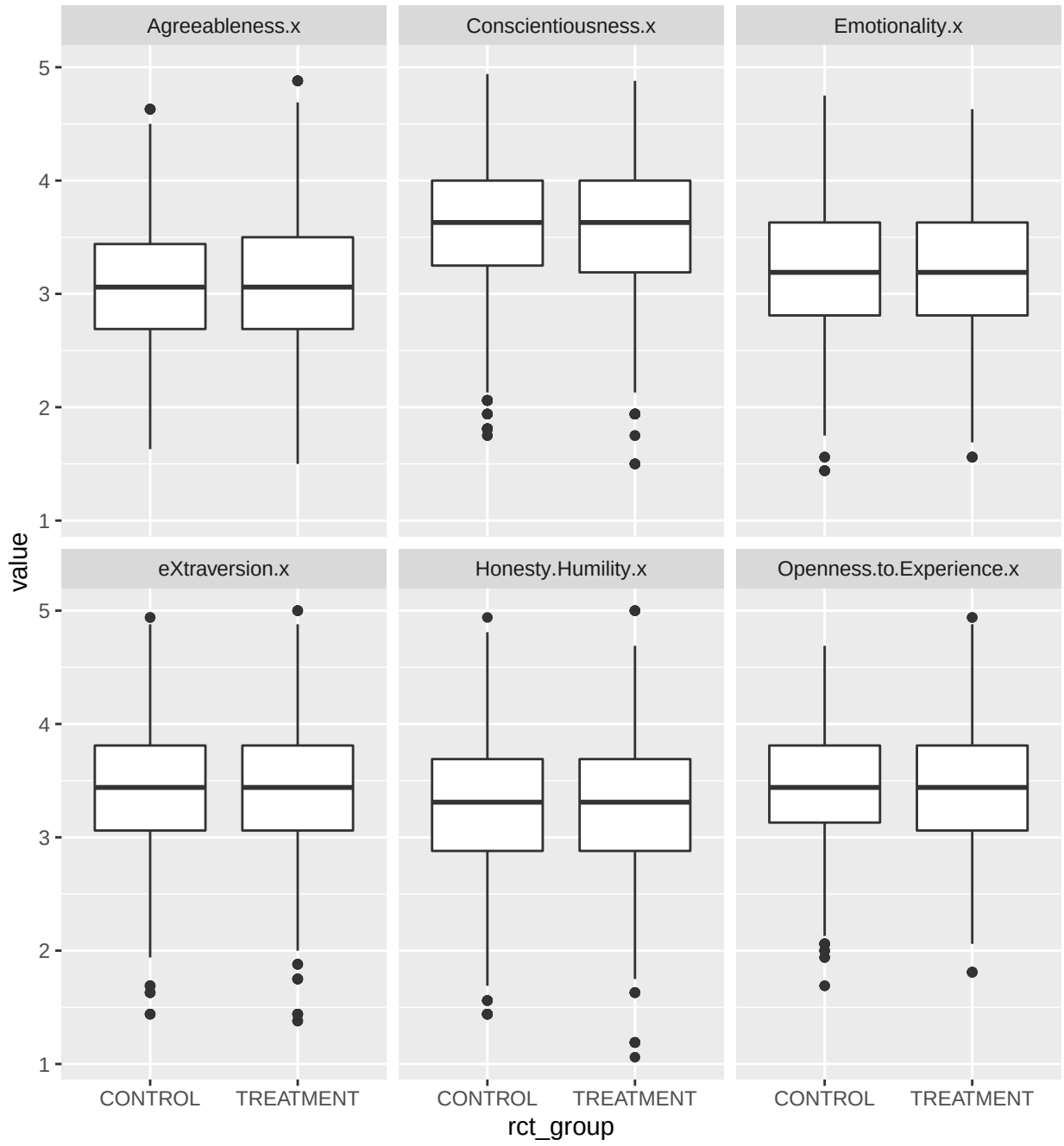
Distribution of Student HEXACO Scores (Across Individual Sessions)

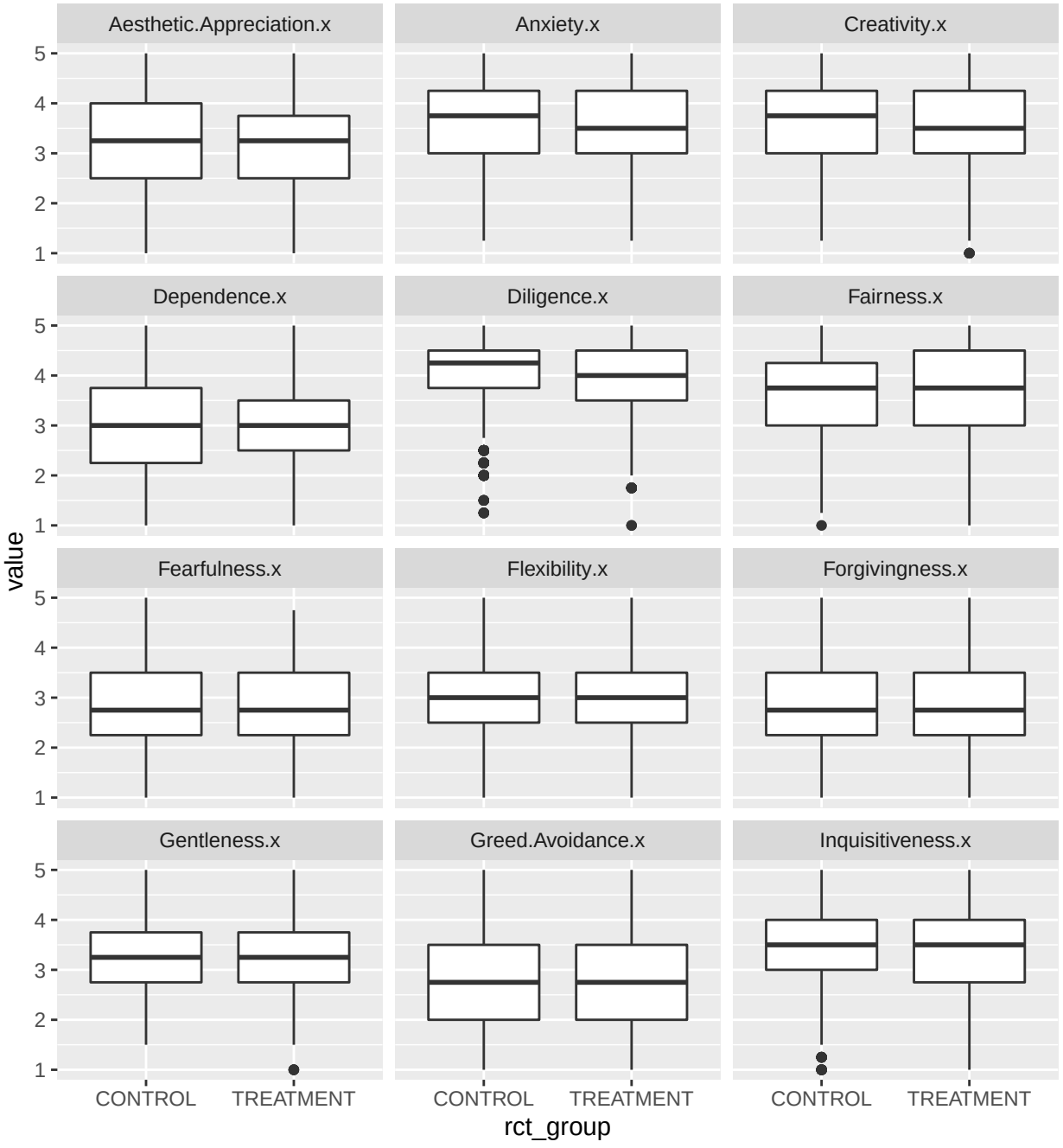


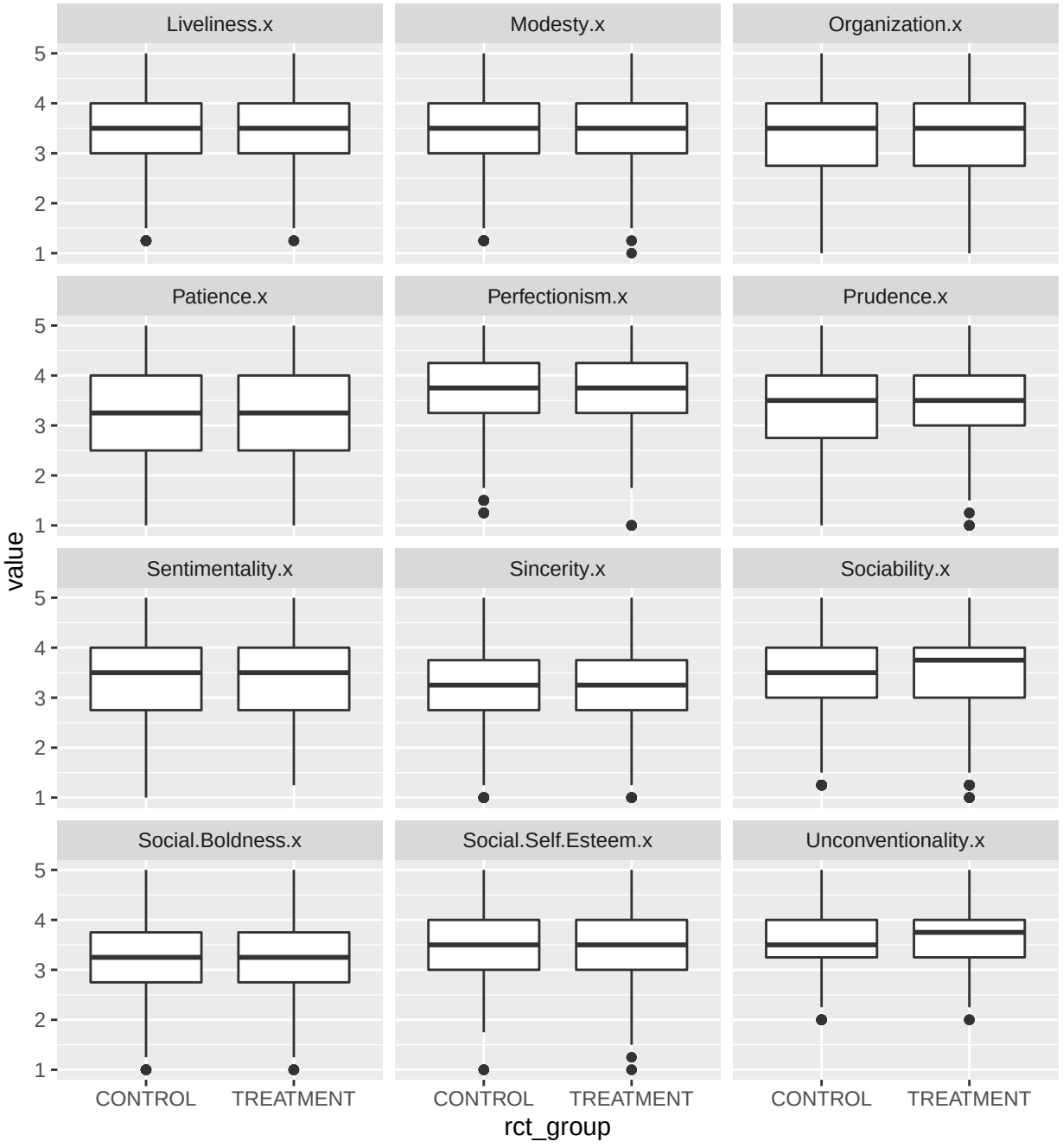

Aesthetic.Appreciation.x
Anxiety.x
Creativity.x
Dependence.x
Diligence.x
Fairness.x
Fearfulness.x
Flexibility.x
Forgivingness.x
Gentleness.x
Greed.Avoidance.x
Inquisitiveness.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group


Liveliness.x
Modesty.x
Organization.x
Patience.x
Perfectionism.x
Prudence.x
Sentimentality.x
Sincerity.x
Sociability.x
Social.Boldness.x
Social.Self.Esteem.x
Unconventionality.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group

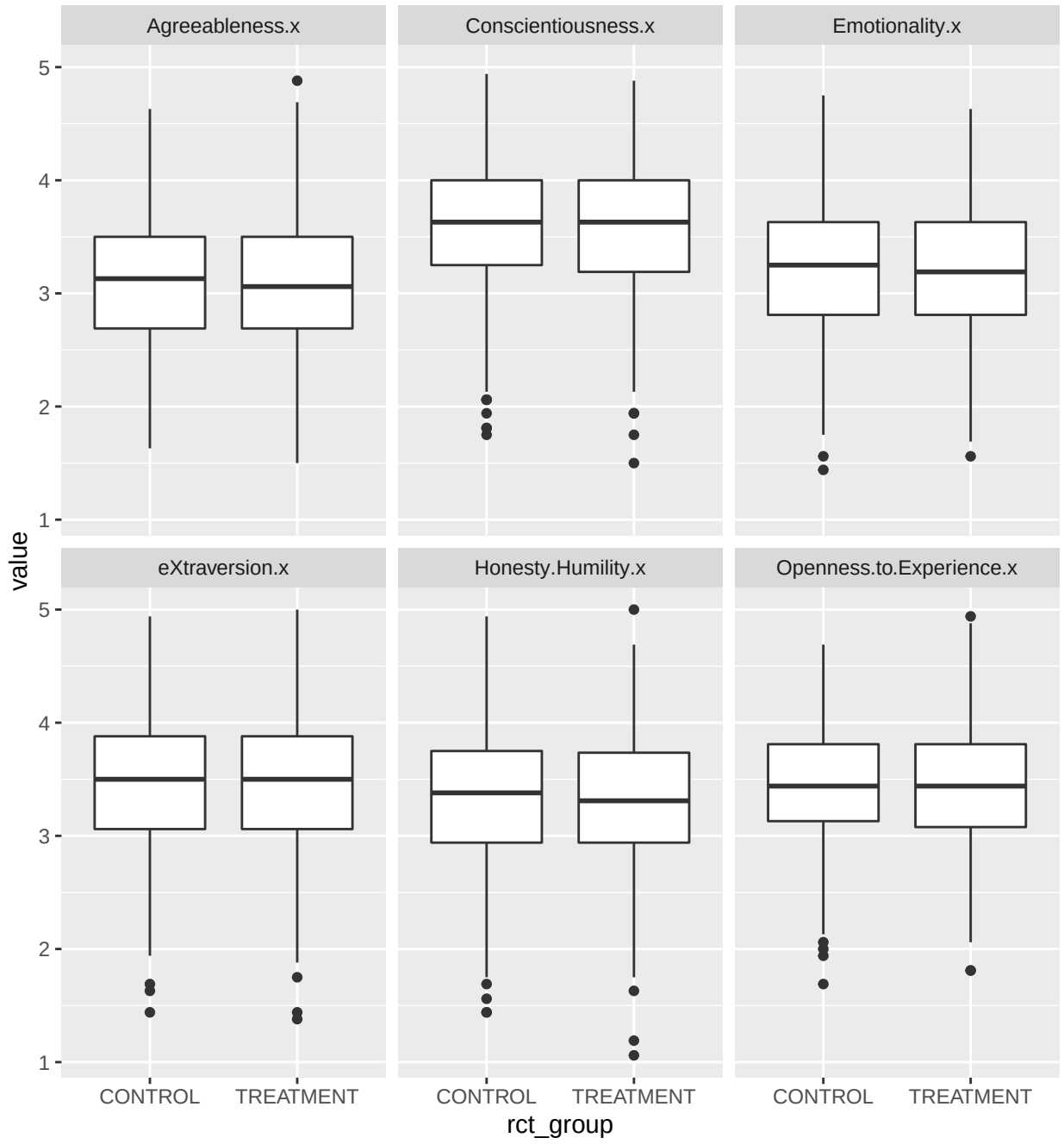
Distribution of Student HEXACO Scores (Across Unique Pairs)

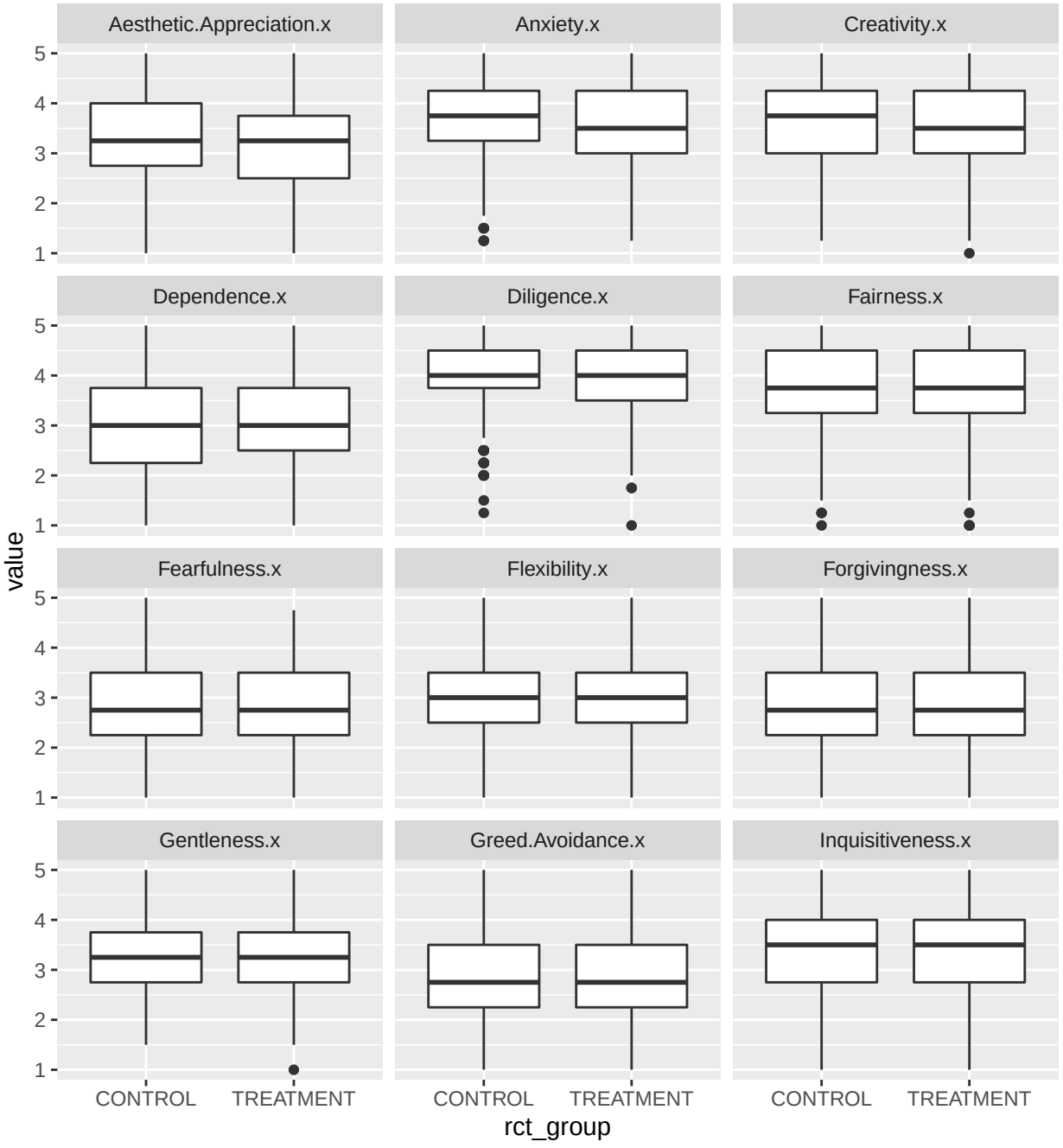


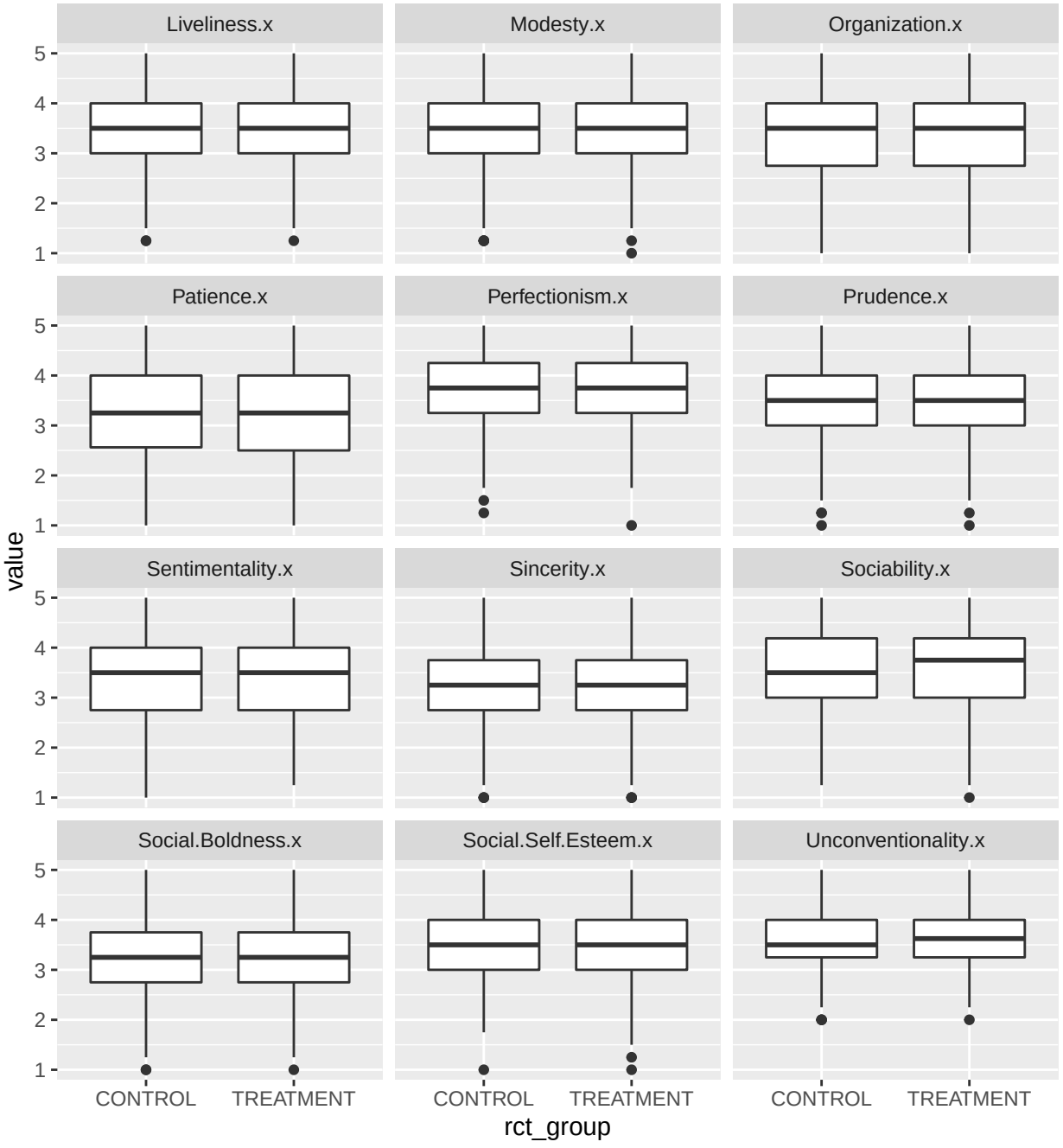

Aesthetic.Appreciation.x
Anxiety.x
Creativity.x
Dependence.x
Diligence.x
Fairness.x
Fearfulness.x
Flexibility.x
Forgivingness.x
Gentleness.x
Greed.Avoidance.x
Inquisitiveness.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group

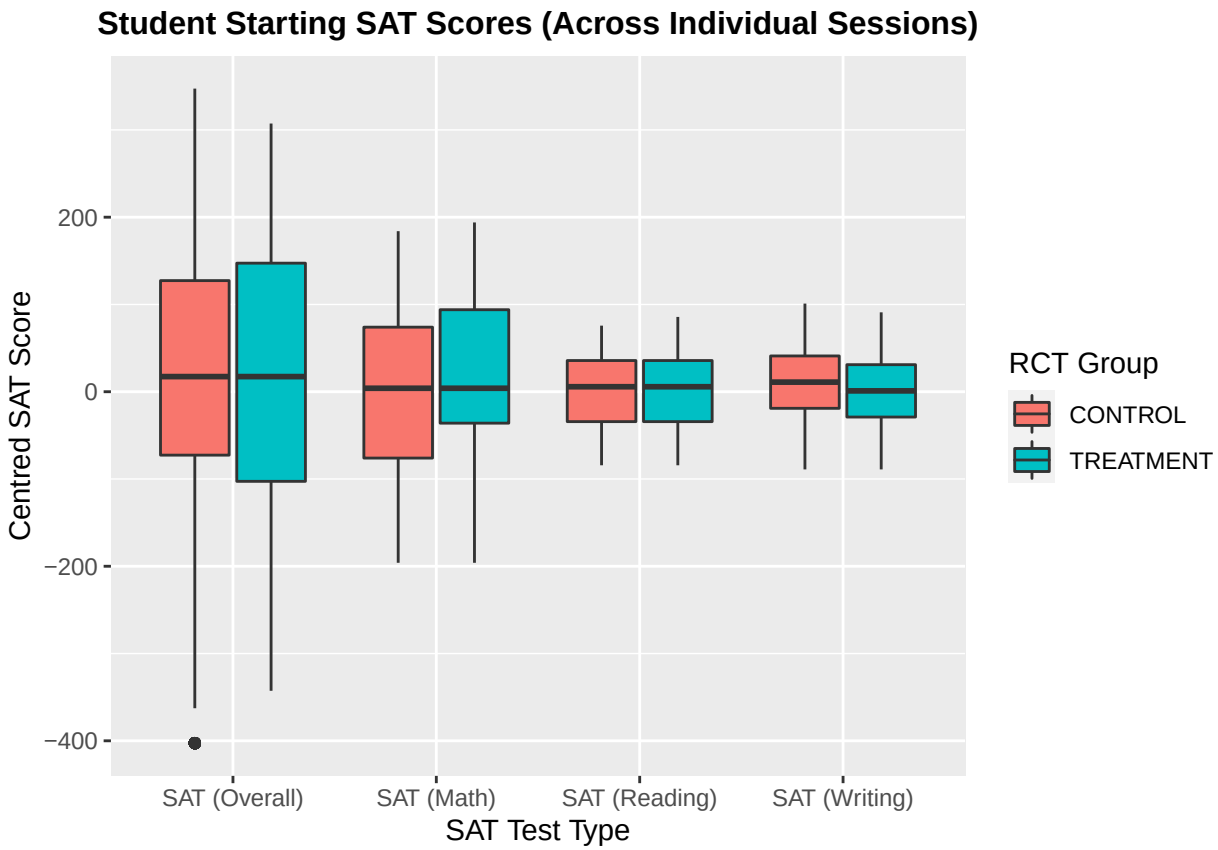

Liveliness.x
Modesty.x
Organization.x
Patience.x
Perfectionism.x
Prudence.x
Sentimentality.x
Sincerity.x
Sociability.x
Social.Boldness.x
Social.Self.Esteem.x
Unconventionality.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group

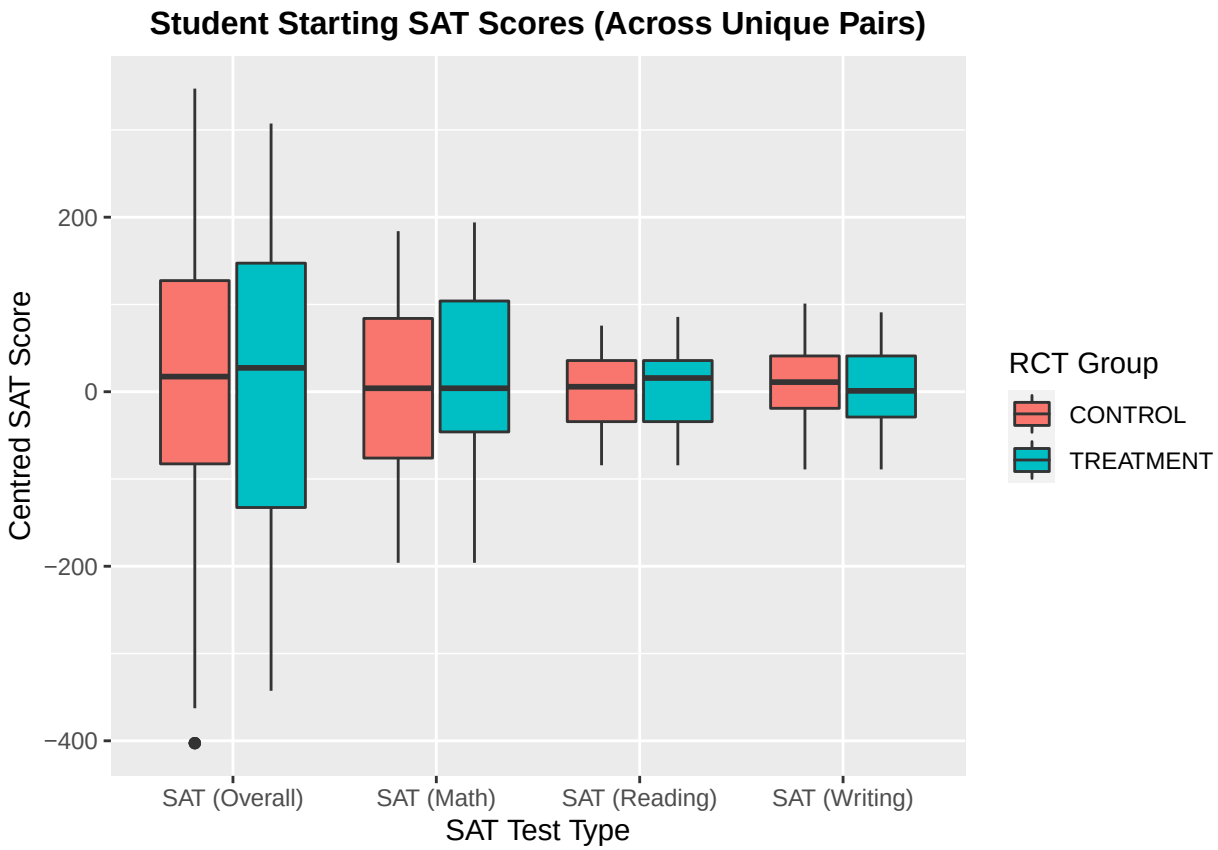
Distribution of Student HEXACO Scores (Across Unique Students)




Aesthetic.Appreciation.x
Anxiety.x
Creativity.x
Dependence.x
Diligence.x
Fairness.x
Fearfulness.x
Flexibility.x
Forgivingness.x
Gentleness.x
Greed.Avoidance.x
Inquisitiveness.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group


Liveliness.x
Modesty.x
Organization.x
Patience.x
Perfectionism.x
Prudence.x
Sentimentality.x
Sincerity.x
Sociability.x
Social.Boldness.x
Social.Self.Esteem.x
Unconventionality.x
value
1
2
3
4
5
CONTROL
TREATMENT
rct_group


Student Starting SAT Scores (Across Individual Sessions)
Centred SAT Score
200
0
−200
−400
SAT (Overall)
SAT (Math)
SAT (Reading)
SAT (Writing)
SAT Test Type
RCT Group
CONTROL
TREATMENT


Student Starting SAT Scores (Across Unique Pairs)
Centred SAT Score
200
0
−200
−400
SAT (Overall)
SAT (Math)
SAT (Reading)
SAT (Writing)
SAT Test Type
RCT Group
CONTROL
TREATMENT

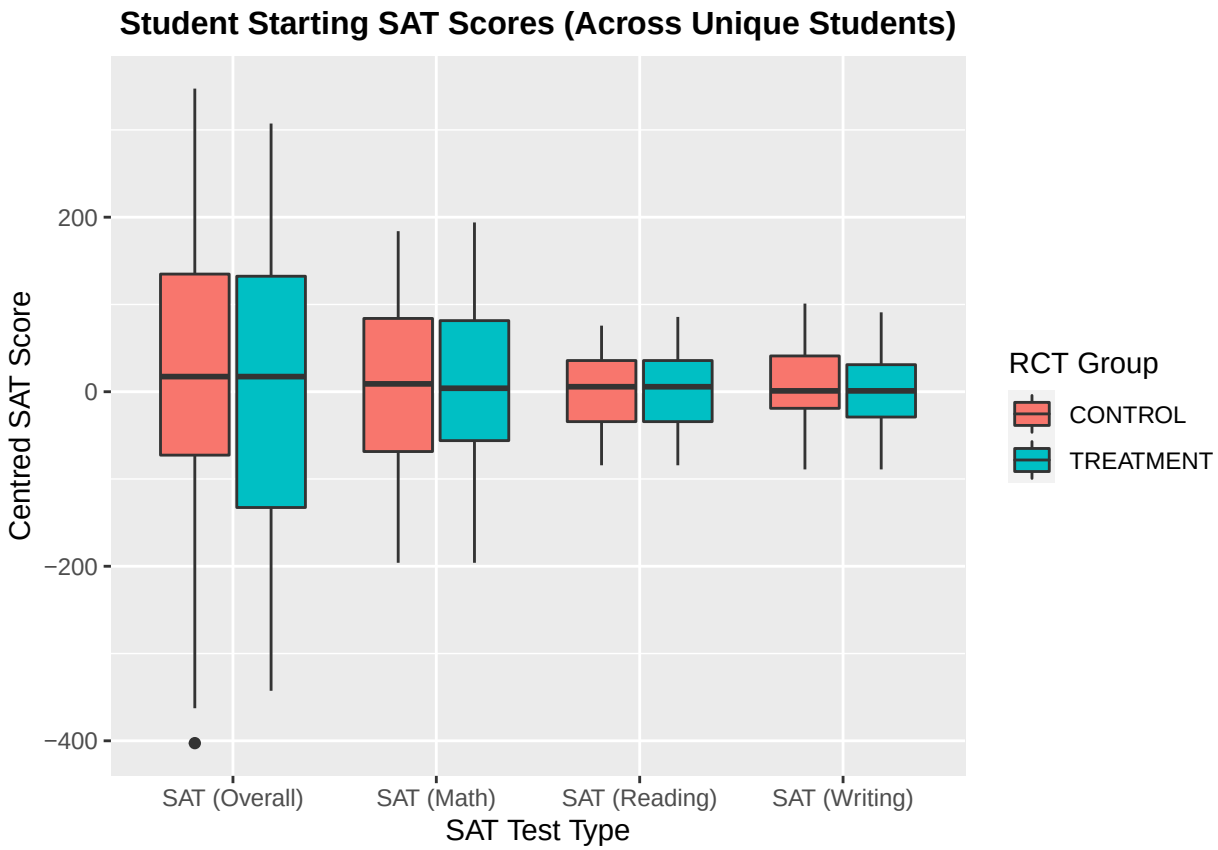

Student Starting SAT Scores (Across Unique Students)
Centred SAT Score
200
0
−200
−400
SAT (Overall)
SAT (Math)
SAT (Reading)
SAT (Writing)
SAT Test Type
RCT Group
CONTROL
TREATMENT

Table C.5: Number of Sessions, Pairs and Tutors by Tutor Country

Tutor Region	# Sessions	# Pairs	# Unique Tutors
Asia	1586	169	33
Australia	3325	353	99
New Zealand	1303	178	72
Other	510	46	15
United Kingdom	5294	517	132
United States	8631	1031	223
NA's	7946	1804	129

Table C.6: Number of Sessions, Pairs and Tutors by Tutor Gender

Tutor Gender	# Sessions	# Pairs	# Unique Tutors
N/A	12102	1767	277
Female	8858	1228	221
Gender diverse/non-binary	132	26	2
Male	7503	1077	203

Tutor age is calculated as at 2 May 2020.

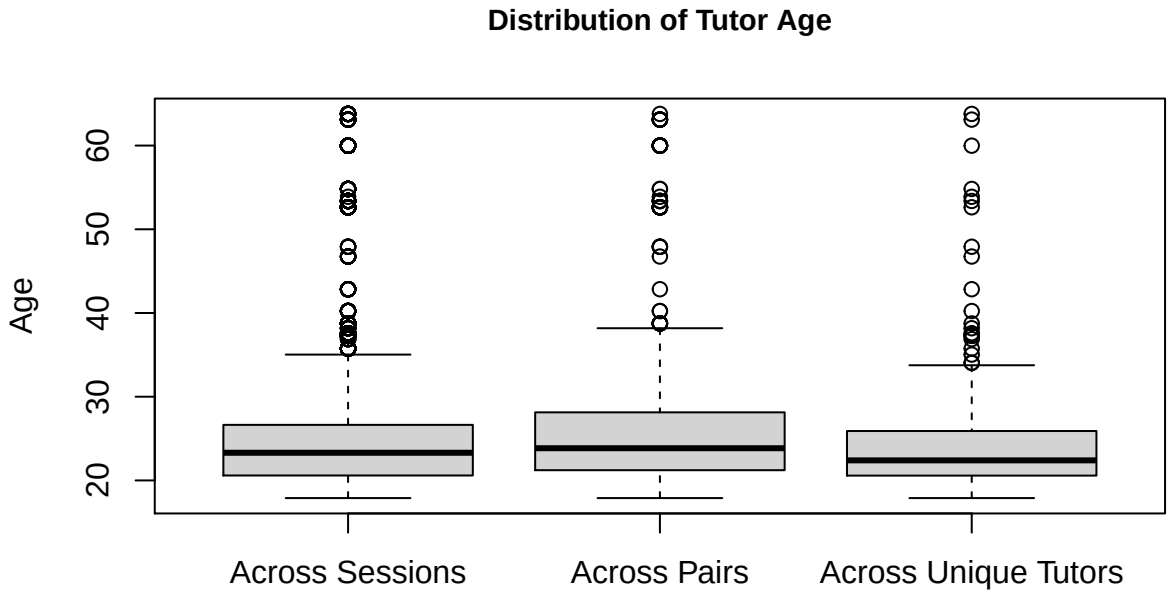
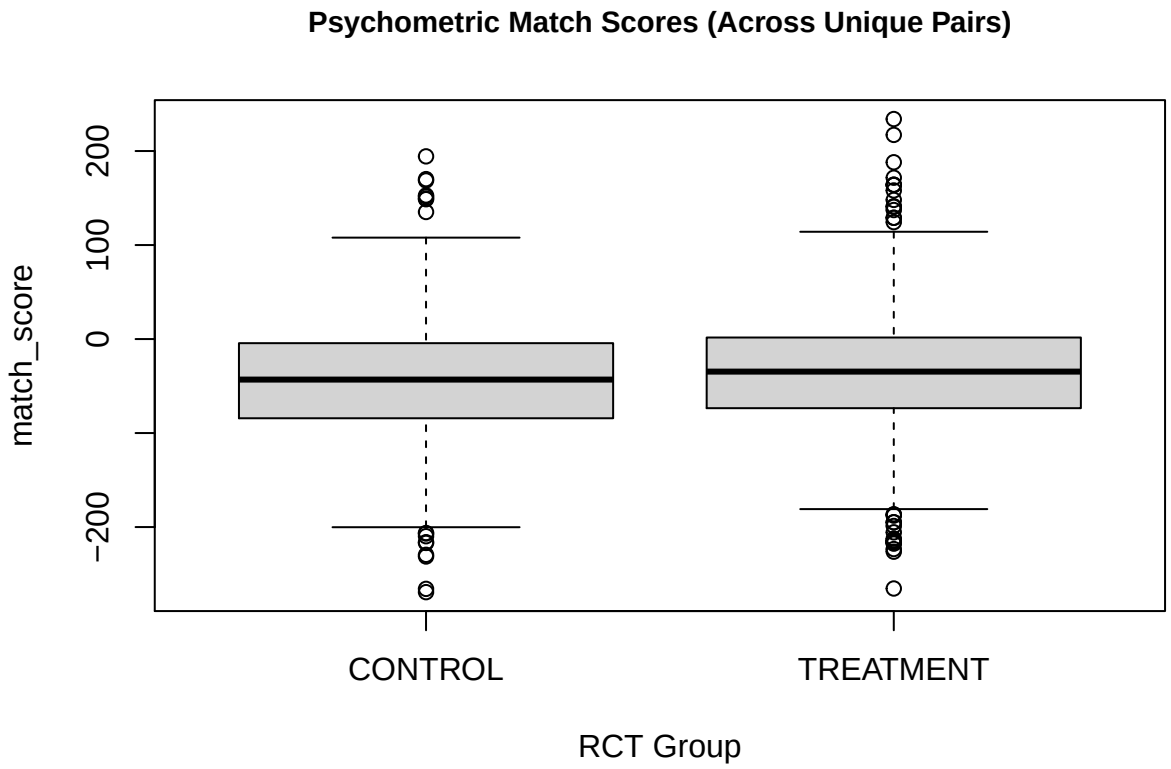
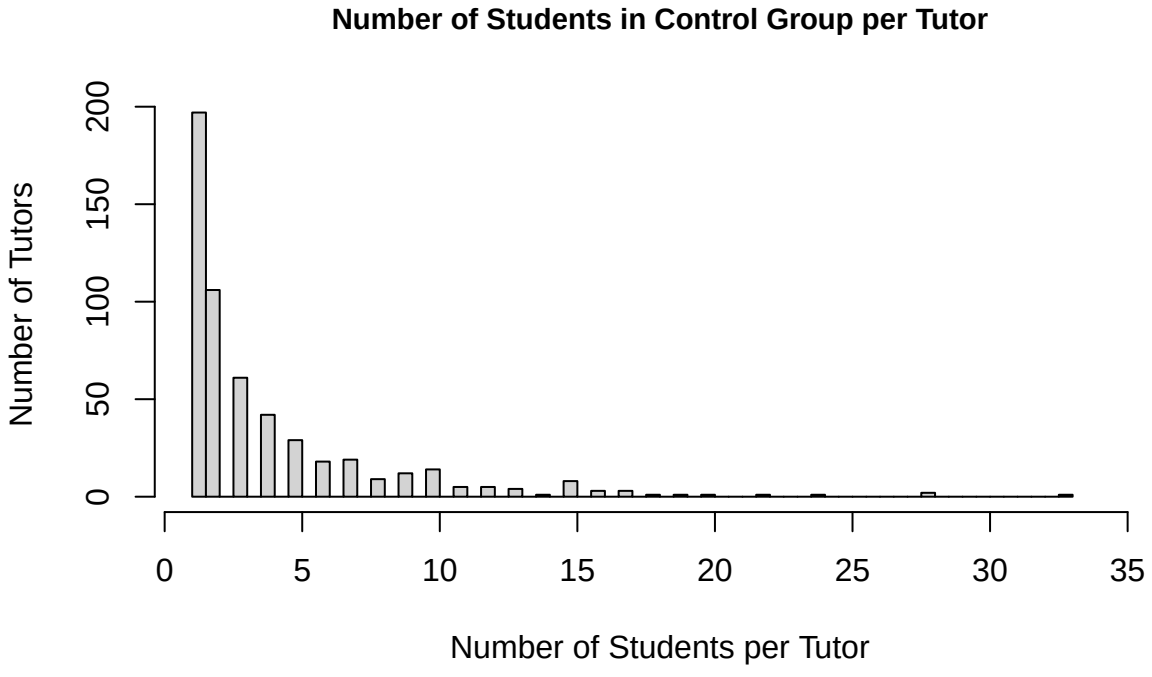


Table C.7: Summary Table of Tutor HEXACO Scores

	Individual Sessions		Unique Pairs		Unique Tutors	
Variable	Mean	S.Dev	Mean	S.Dev	Mean	S.Dev
Aesthetic Appreciation	3.66	0.85	3.70	0.85	3.68	0.80
Agreeableness	3.25	0.54	3.25	0.55	3.23	0.55
Anxiety	3.56	0.85	3.48	0.86	3.46	0.84
Conscientiousness	3.94	0.48	3.92	0.48	3.89	0.49
Creativity	3.75	0.76	3.77	0.76	3.78	0.78
Dependence	3.21	0.84	3.16	0.86	3.22	0.81
Diligence	4.36	0.52	4.34	0.52	4.31	0.54
Emotionality	3.30	0.58	3.24	0.60	3.26	0.56
eXtraversion	3.76	0.59	3.75	0.58	3.71	0.58
Fairness	4.03	0.83	4.00	0.80	3.93	0.82
Fearfulness	2.81	0.79	2.76	0.79	2.83	0.77
Flexibility	3.24	0.75	3.18	0.76	3.13	0.73
Forgivingness	2.96	0.78	3.00	0.78	2.96	0.75
Gentleness	3.29	0.70	3.30	0.71	3.28	0.71
Greed Avoidance	3.20	1.00	3.20	0.96	3.23	0.91
Honesty Humility	3.54	0.59	3.53	0.57	3.53	0.56
Inquisitiveness	3.91	0.74	3.91	0.74	3.88	0.73
Liveliness	3.72	0.75	3.71	0.75	3.66	0.77
Modesty	3.58	0.76	3.60	0.75	3.64	0.75
Openness to Experience	3.80	0.54	3.82	0.54	3.80	0.53
Organization	3.85	0.82	3.86	0.79	3.82	0.82
Patience	3.49	0.83	3.52	0.82	3.52	0.77
Perfectionism	3.89	0.68	3.87	0.67	3.84	0.65
Prudence	3.65	0.62	3.62	0.64	3.60	0.65
Sentimentality	3.61	0.79	3.57	0.77	3.53	0.76
Sincerity	3.34	0.79	3.29	0.79	3.30	0.78
Sociability	3.75	0.76	3.74	0.74	3.71	0.75
Social Boldness	3.55	0.83	3.55	0.82	3.51	0.81
Social Self Esteem	4.03	0.61	3.98	0.65	3.94	0.65
Unconventionality	3.88	0.60	3.88	0.60	3.86	0.60


Psychometric Match Scores (Across Unique Pairs)
match_score
200
100
0
−200
CONTROL
TREATMENT
RCT Group


Number of Students in Control Group per Tutor
Number of Tutors
0
50
100
150
200
0
5
10
15
20
25
30
35
Number of Students per Tutor

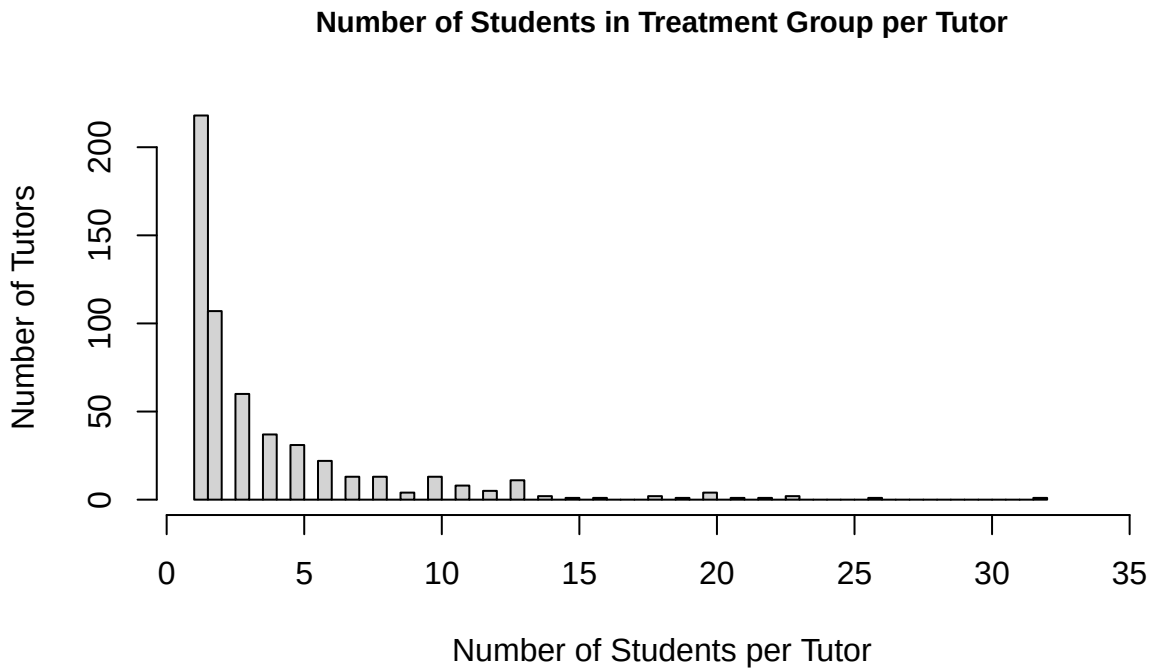


Table C.8: Average Session Rating by Student Country

	Average Rating	
Student Region	Across Sessions	Across Pairs
Australia	4.69	4.66
East / South-East / South Asia	4.75	4.64
Eastern Europe & Central Asia	4.85	4.73
Latin America	4.80	4.74
Middle East & Africa	4.65	4.65
New Zealand	4.77	4.70
North America	4.70	4.64
Singapore	4.67	4.64
Western Europe	4.65	4.62
NA	4.62	4.62

Table C.9: Number of Unique Pairs by Student Gender (with Starting and Ending SAT Scores)

	CONTROL	TREATMENT
N/A	18	9
Female	4	6
Male	3	6

Table C.10: Number of Students by Country (Control vs Treatment Group, with Valid Starting and Ending Overall SAT Scores)

	CONTROL	TREATMENT
Australia	5	3
East / South-East / South Asia	3	3
Eastern Europe & Central Asia	2	4
Latin America	1	3
New Zealand	4	2
North America	2	1
Singapore	0	3
Western Europe	8	2

Distribution of Psychometric Match Scores (Control vs Treatment, Across Unique Pairs Containing Students with Valid Starting and Ending Overall SAT Scores)

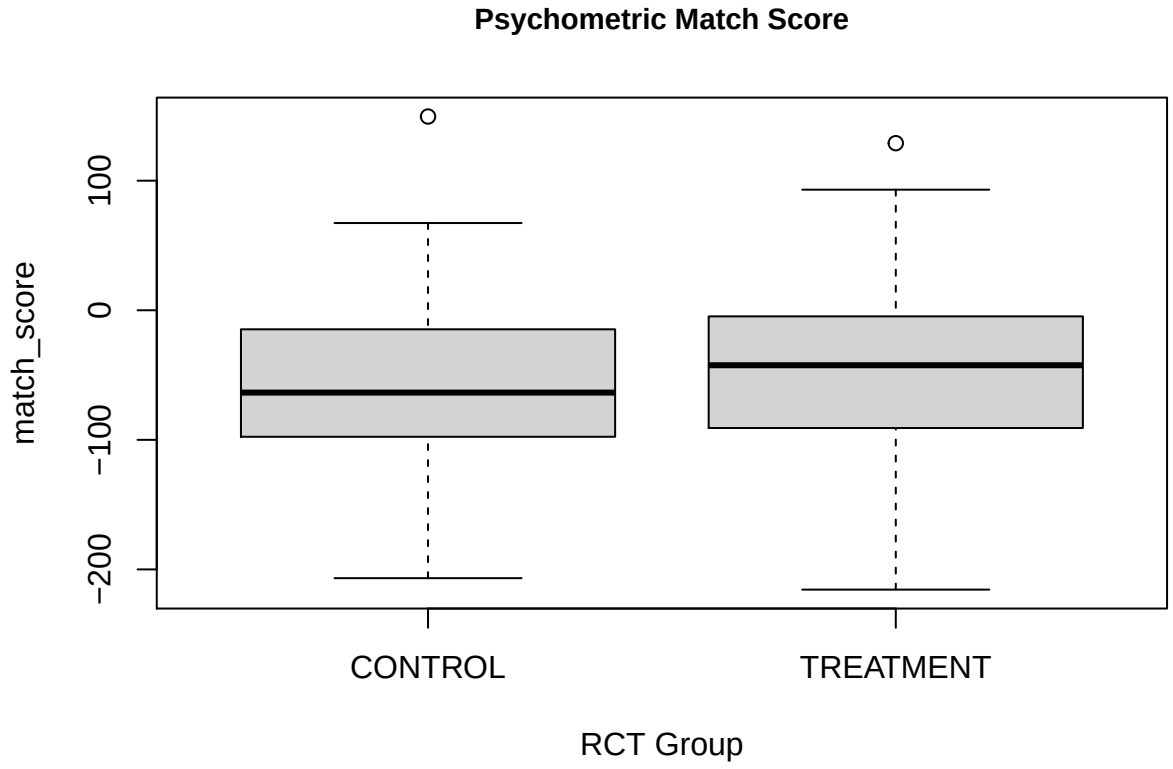


Table C.11: Number of Students by RCT Group and SAT Group

	Initial & Final Tests	Initial Test Only	No Test Result
CONTROL	25	125	393
TREATMENT	21	126	405

Table C.12: Number of Students by Gender and SAT Group

	Initial & Final Tests	Initial Test Only	No Test Result
N/A	27	132	484
Female	10	59	163
Male	9	60	151

Table C.13: Number of Students by Country and SAT Group

	Initial & Final Tests	Initial Test Only	No Test Result
Australia	8	31	146
East / South-East / South Asia	6	44	103
Eastern Europe & Central Asia	6	25	75
Latin America	4	19	42
Middle East & Africa	0	22	41
New Zealand	6	27	94
North America	3	29	80
Singapore	3	12	84
Western Europe	10	36	82
NA	0	6	51

Table C.14: Descriptive Statistics - Tutors with < 5 Students vs 5+ Students

Students per Tutor	Number of Tutors
At least 5	257
Fewer than 5	446

Table C.15: Number of Tutors by Tutor Gender

	At least 5	Fewer than 5
N/A	113	164
Female	77	144
Gender diverse/non-binary	1	1
Male	66	137

Table C.16: Pearson's Chi-Squared Test of Independence on Proportion of Tutors by Gender (< 5 Students vs 5+ Students)

Test statistic	df	P value
12	9	0.2133

Table C.17: Number of Tutors by Region (< 5 Students vs 5+ Students)

	At least 5	Fewer than 5
Asia	14	19
Australia	29	70
New Zealand	8	64
Other	2	13
United Kingdom	41	91
United States	69	154
NA	94	35

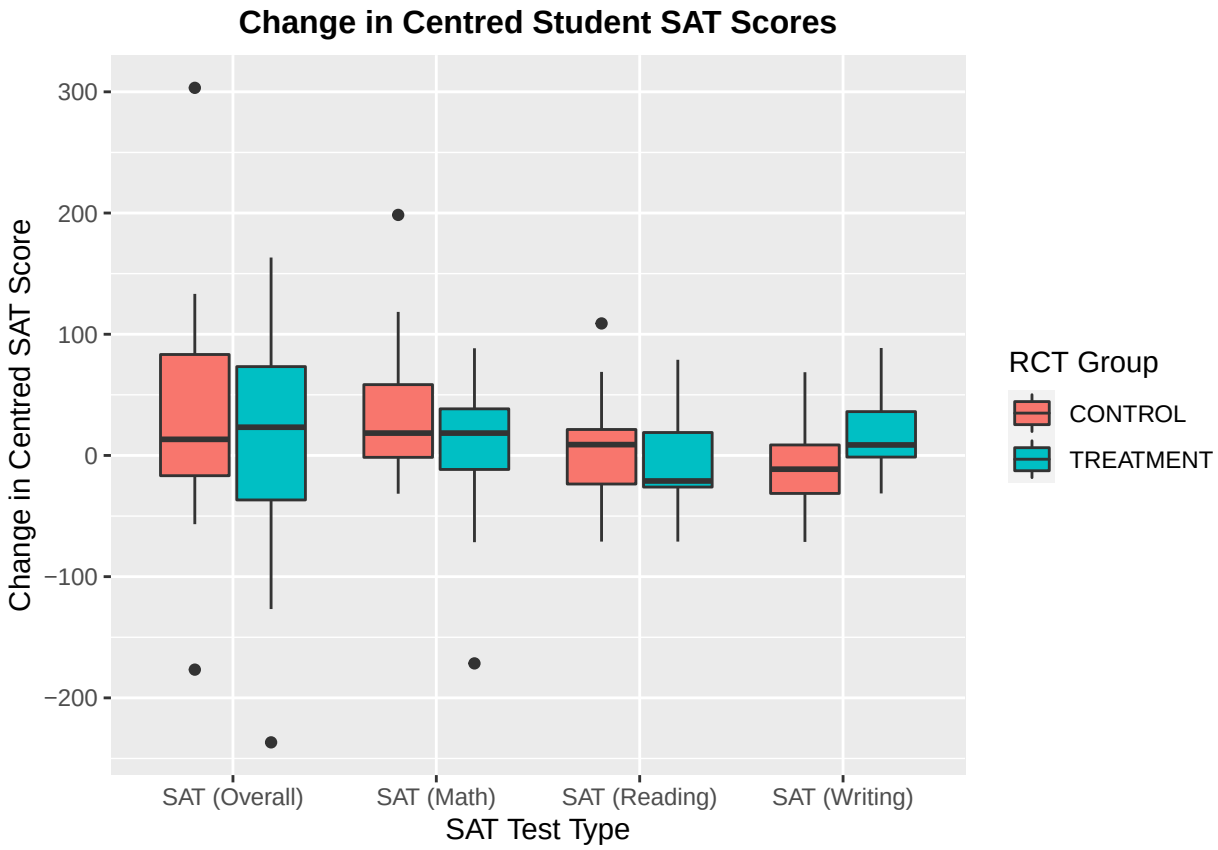
Table C.18: Pearson's Chi-Squared Test of Independence on Proportion of Tutors by Region (< 5 Students vs 5+ Students)

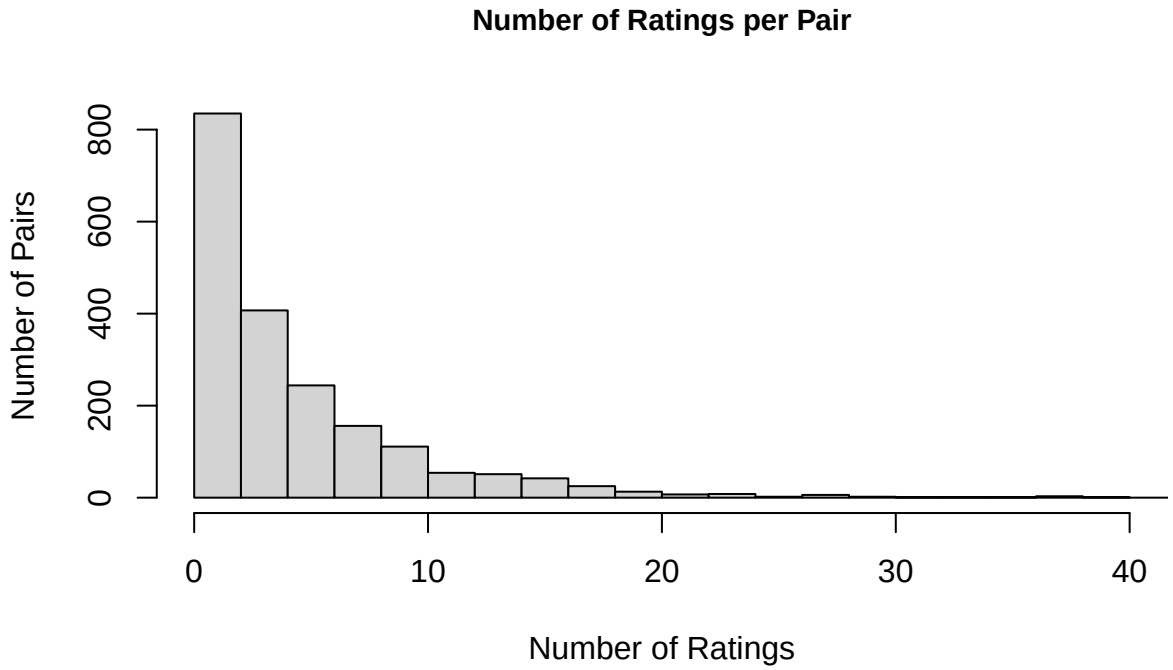
Test statistic	df	P value
42	36	0.227

Table C.19: Tutor HEXACO by Number of Students Taught

	Group Means			
Variable	At least 5	Fewer than 5	p value	t value
Aesthetic Appreciation	3.71	3.67	0.66	-0.44
Agreeableness	3.24	3.22	0.61	-0.50
Anxiety	3.48	3.45	0.66	-0.44
Conscientiousness	3.92	3.88	0.47	-0.72
Creativity	3.84	3.76	0.23	-1.21
Dependence	3.23	3.22	0.93	-0.09
Diligence	4.36	4.29	0.14	-1.47
Emotionality	3.26	3.26	0.96	-0.05
eXtraversion	3.77	3.68	0.09	-1.72
Fairness	4.07	3.87	0.01	-2.79
Fearfulness	2.74	2.86	0.08	1.73
Flexibility	3.18	3.12	0.38	-0.89
Forgivingness	3.03	2.93	0.14	-1.50
Gentleness	3.32	3.27	0.51	-0.66
Greed Avoidance	3.21	3.23	0.80	0.25
Honesty Humility	3.57	3.51	0.25	-1.15
Inquisitiveness	3.99	3.84	0.02	-2.26

Liveliness	3.75	3.63	0.08	-1.73
Modesty	3.63	3.64	0.90	0.12
Openness to Experience	3.85	3.78	0.15	-1.43
Organization	3.84	3.81	0.69	-0.40
Patience	3.45	3.55	0.16	1.42
Perfectionism	3.86	3.84	0.70	-0.38
Prudence	3.60	3.60	0.94	-0.08
Sentimentality	3.60	3.51	0.17	-1.37
Sincerity	3.35	3.28	0.34	-0.96
Sociability	3.81	3.67	0.03	-2.17
Social Boldness	3.55	3.50	0.49	-0.69
Social Self Esteem	3.97	3.92	0.38	-0.88
Unconventionality	3.87	3.85	0.78	-0.28




Number of Ratings per Pair
Number of Pairs
0
200
400
600
800
0
10
20
30
40
Number of Ratings

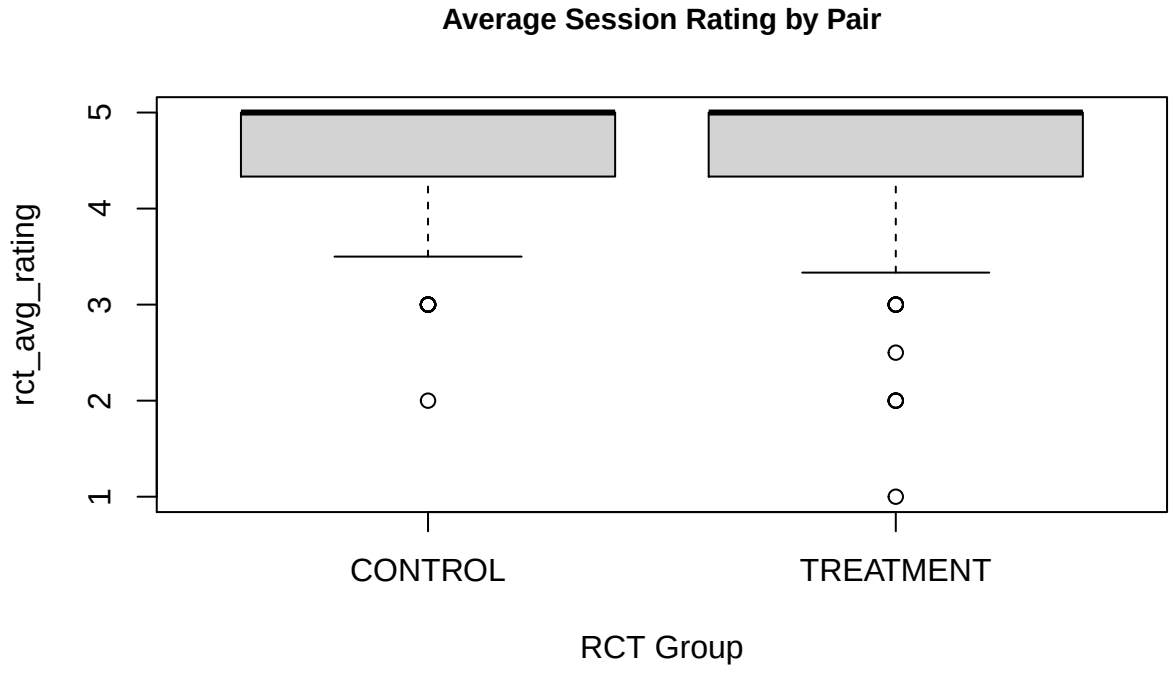

Average Session Rating by Pair
rct_avg_rating
1
2
3
4
5
CONTROL
TREATMENT
RCT Group

Table C.20: Average Session Rating (Across Unique Pairs) by Student Country

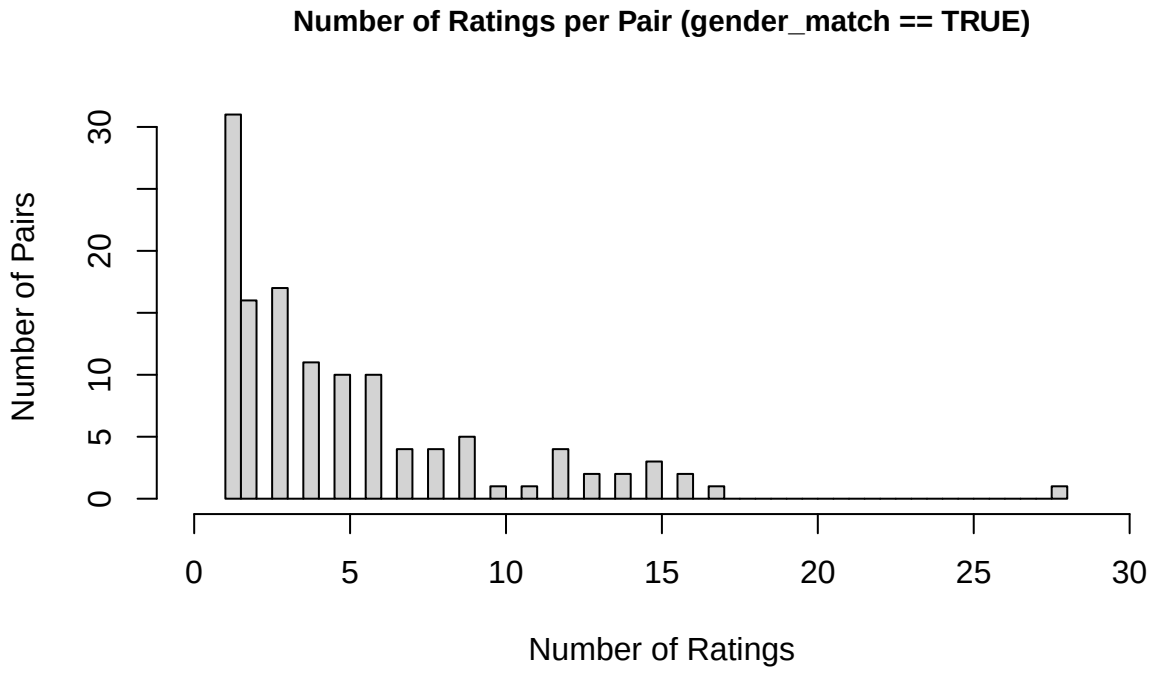
Group	Country	Number of Pairs			NA
		With Ratings	Without Ratings	Number of Ratings	
control	Australia	4.69	185	177	923
treatment	Australia	4.62	136	123	662
control	East / South-East / South Asia	4.70	116	158	521
treatment	East / South-East / South Asia	4.56	105	176	635
control	Eastern Europe & Central Asia	4.82	157	121	1120
treatment	Eastern Europe & Central Asia	4.62	140	107	981
control	Latin America	4.60	31	63	111
treatment	Latin America	4.82	56	73	208
control	Middle East & Africa	4.65	59	60	357
treatment	Middle East & Africa	4.65	78	71	378
control	New Zealand	4.73	114	95	487
treatment	New Zealand	4.68	146	131	688
control	North America	4.60	82	104	406
treatment	North America	4.68	110	127	421
control	Singapore	4.59	80	98	341
treatment	Singapore	4.71	67	101	294
control	Western Europe	4.54	127	124	483
treatment	Western Europe	4.74	84	105	331

Gender Match and Session Rating

Here, we look at solely the RCT Control group in case gender is correlated with the matching algorithm.

Table C.21: Gender and Session Rating - Overview

	Gender Match		
	FALSE	TRUE	N/A
Number of Pairs	212.000	224.000	1622.000
Average Number of Ratings	5.536	4.944	4.896
Average Number of Sessions	7.335	7.308	7.012
Average Pair Rating	4.538	4.707	4.682
Average Session Rating	4.721	4.728	4.749



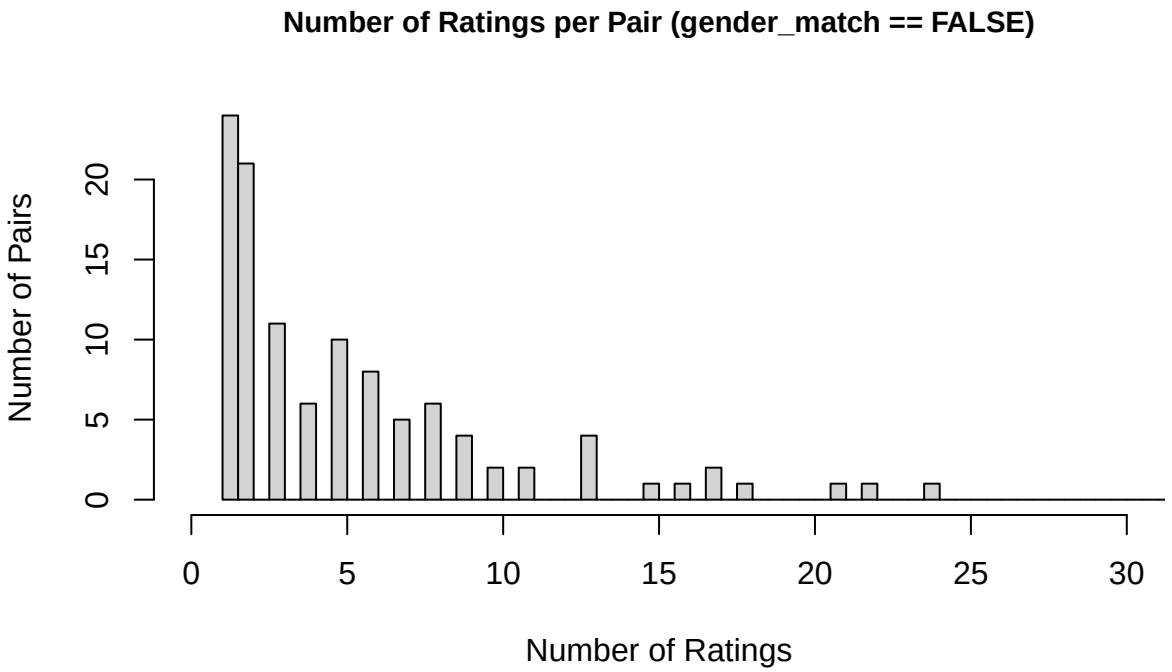


Table C.22: Gender and Session Rating - (Pairs with 2 or Fewer vs 3 or More Ratings

	2 or Fewer Ratings			3 or More Ratings		
	Gender Match					
	FALSE	TRUE	N/A	FALSE	TRUE	N/A
Number of Pairs	45.000	47.000	1375.000	67.000	78.000	1496.000
Average Number of Ratings	1.467	1.340	1.409	8.269	7.115	7.437
Average Number of Sessions	3.644	3.553	3.677	15.164	13.538	14.392
Average Pair Rating	4.333	4.670	4.622	4.675	4.730	4.725

Table C.23: 2-Sample T-Test on Number of Ratings per Pair (Grouped by Gender Match)

	Group Mean by Gender Match				
	Not Matched	Matched	T Statistic	P Value	df
Ratings per Pair	5.5	4.9	0.869	0.386	212
Sessions per Pair	7.3	7.3	0.034	0.973	429

Table C.24: OLS Regression - Session Rating vs Centred SAT Score Improvement (Overall and Math)

	Dependent variable:			
	Change in Overall SAT		Change in SAT Math	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
rct_avg_rating	19.239	6.742	16.095	11.955
	(17.106)	(18.892)	(11.170)	(12.488)
Constant	−82.292	−25.367	−58.787	−40.257
	(81.048)	(90.159)	(52.912)	(59.603)
Observations	122	122	113	113
R^2	0.010	0.001	0.018	0.008
Adjusted R^2	0.002	−0.007	0.010	−0.001
Residual Std. Error	84.098	122.657	52.553	77.996
F Statistic	1.265	0.127	2.076	0.916
Note:				*p<0.1; **p<0.05; ***p<0.01

Models: (1) SAT Overall Improvement - Unweighted model (2) SAT Overall Improvement - Weighted by square root of number of ratings for each pair (3) SAT Math Improvement - Unweighted model (4) SAT Math Improvement - Weighted by square root of number of ratings for each pair

Table C.25: OLS Regression - Session Rating vs Centred SAT Score Improvement (Reading and Writing)

	Dependent variable:			
	Change in SAT Reading		Change in SAT Writing	
	b (SE)	b (SE)	b (SE)	b (SE)
	(1)	(2)	(3)	(4)
rct_avg_rating	0.118	−3.473	−11.174	−13.496*
	(7.119)	(7.950)	(7.079)	(7.639)
Constant	−6.306	9.019	52.647	61.732*
	(33.770)	(37.994)	(33.531)	(36.364)
Observations	124	124	108	108
R^2	0.00000	0.002	0.023	0.029
Adjusted R^2	−0.008	−0.007	0.014	0.019
Residual Std. Error	36.495	52.759	33.121	47.743
F Statistic	0.0003	0.191	2.491	3.121*
Note:				*p<0.1; **p<0.05; ***p<0.01

Models: (1) SAT Reading Improvement - Unweighted model (2) SAT Reading Improvement - Weighted by square root of number of ratings for each pair (3) SAT Writing Improvement - Unweighted model (4) SAT Writing Improvement - Weighted by square root of number of ratings for each pair

D

Citations

Baker, Beryle I., Pearl Henry, and Newburn Reynolds. "Reflections on an NCATE Study: The Recruitment and Retention of Males and Minorities in National Council of Accreditation Teacher Education (NCATE) Institutions: The Role of the Two-Year College in Teacher Education." (1998).

Banerjee, Neena. "Student–teacher ethno-racial matching and reading ability group placement in early grades." *Education and Urban Society* 51.3 (2019): 395-422.

Banks, James A. "Multicultural Education, Transformative Knowledge and Action: Historical and Contemporary Perspectives. Multicultural Education Series." (1996).

Barbour, Michael, and Dennis Mulcahy. "Student performance in virtual schooling: Looking beyond the numbers." (2009).

Basham, James D., et al. "The scaled arrival of K-12 online education: Emerging realities and implications for the future of education." *Journal of Education* 193.2 (2013): 51-60.

Bettinger, Eric, et al. "Increasing perseverance in math: Evidence from a field experiment in Norway." *Journal of Economic Behavior & Organization* 146 (2018): 1-15.

Borup, Jered, Charles R. Graham, and Randall S. Davies. "The nature of adolescent learner interaction in a virtual high school setting." *Journal of Computer Assisted Learning* 29.2 (2013): 153-167.

Brown-Jeffy, Shelly. "School effects: Examining the race gap in mathematics achievement." *Journal of African American Studies* 13.4 (2009): 388-405.

Caine, Akia D., Jacob Schwartzman, and Anastasia Kunac. "Speed dating for mentors: a novel approach to mentor/mentee pairing in surgical residency." *Journal of Surgical Research* 214 (2017): 57-61.

Campbell, Toni A., and David E. Campbell. "Faculty/student mentor program: Effects on academic performance and retention." *Research in higher education* 38.6 (1997): 727-742.

Casteel, Clifton A. "Teacher–student interactions and race in integrated classrooms." *The Journal of Educational Research* 92.2 (1998): 115-120.

Cavanaugh, Cathy S., Michael K. Barbour, and Tom Clark. "Research and practice in K-12 online learning: A review of open access literature." *The International Review of Research in Open and Distributed Learning* 10.1 (2009).

Cavanaugh, Cathy, et al. "The effects of distance education on k-12 student outcomes: A meta-analysis." *Learning Point Associates/North Central Regional Educational Laboratory (NCREL)* (2004).

Cho, Insook. "The effect of teacher–student gender matching: Evidence from OECD countries." *Economics of Education Review* 31.3 (2012): 54-67

Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. "Teacher-student matching and the assessment of teacher effectiveness." *Journal of human Resources* 41.4 (2006): 778-820.

Cosentino de Cohen, Clemencia, Nicole Deterding, and Beatriz Chu Clewell. "Who's Left Behind? Immigrant Children in High and Low LEP Schools." *Urban*

Institute (NJ3) (2005).

Coutts, Sharona. "At University of Phoenix, Allegations of Enrollment Abuses Persist." ProPublica. org 3 (2009).

Crosnoe, Robert, Monica Kirkpatrick Johnson, and Glen H. Elder Jr. "Intergenerational bonding in school: The behavioral and contextual correlates of student-teacher relationships." *Sociology of education* 77.1 (2004): 60-81.

Curtis, Heidi, and Loredana Werth. "Fostering student success and engagement in a K-12 online school." *Journal of Online Learning Research* 1.2 (2015): 163-190.

Dede, Chris, John Richards, and Bror Saxberg, eds. Learning engineering for online education: Theoretical contexts and design-based examples. Routledge, 2018.

Dee, Thomas S. "A teacher like me: Does race, ethnicity, or gender matter?." *American Economic Review* 95.2 (2005): 158-165.

Dee, Thomas S. "The race connection: Are teachers more effective with students who share their ethnicity?." *Education Next* 4.2 (2004): 52-60.

Department of Industry, Innovation, Climate Change, Scientific Research, and Tertiary Education, 2013.

Driscoll, Adam, et al. "Can online courses deliver in-class results? A comparison of student performance and satisfaction in an online versus a face-to-face introductory sociology course." *Teaching Sociology* 40.4 (2012): 312-331.

Drysdale, Jeffery, Charles Graham, and Jered Borup. "An online high school "shepherding" program: Teacher roles and experiences mentoring online students." *Journal of Technology and Teacher Education* 22.1 (2014): 9-32.

Easton-Brooks, Lewis & Zhang 2010 - Easton-Brooks, Donald, C. Lewis, and Yubo Zhang. "Ethnic-matching: The influence of African American teachers on the reading scores of African American students." *National Journal of Urban Education & Practice* 3.1 (2010): 230-243.

Eddy, Colleen M., and Donald Easton-Brooks. "Ethnic matching, school placement, and mathematics achievement of African American students from

kindergarten through fifth grade." *Urban Education* 46.6 (2011): 1280-1299.

Egalite, Anna J., Brian Kisida, and Marcus A. Winters. "Representation in the classroom: The effect of own-race teachers on student achievement." *Economics of Education Review* 45 (2015): 44-52.

El Said, Ghada Refaat. "Understanding how learners use massive open online courses and why they drop out: Thematic analysis of an interview study in a developing country." *Journal of Educational Computing Research* 55.5 (2017): 724-752.

Ellis, Kathleen. "Cyber charter schools: Evolution, issues, and opportunities in funding and localized oversight." *Educational Horizons* 86.3 (2008): 142-152.

Farrington, David P., et al. "The Maryland Scientific Methods Scale." Evidence-based crime prevention. Routledge, 2003. 27-35.

Ferguson, Ronald F. "Teachers' perceptions and expectations and the Black-White test score gap." *Urban education* 38.4 (2003): 460-507.

Fernandez, Heidi, et al. "Students with special health care needs in K-12 virtual schools." *Journal of Educational Technology & Society* 19.1 (2016): 67-75.

Ferrari, Rossella. "Writing narrative style literature reviews." *Medical Writing* 24.4 (2015): 230-235.

Fillmore, Ian. "Price discrimination and public policy in the US college market." Employment Research Newsletter 23.2 (2016): 2.

Foster, Michele. *Black teachers on teaching.* New Press, 500 Fifth Avenue, New York, NY 10110, 1997.

Fulton, Kathleen. "Preserving Principles of Public Education in an Online World: What Policymakers Should Be Asking about Virtual Schools (Washington, DC, April 19, 2002)." (2002).

Gee, James Paul. Situated language and learning: A critique of traditional schooling. Psychology Press, 2004.

Gifford III, William M. *Online High School Student Achievement on State-Issued Standardized Tests: A Case Study.* Diss. Northcentral University, 2017.

Grossman, Sanford J., and Oliver D. Hart. "An analysis of the principal-agent problem." Foundations of Insurance Economics. Springer, Dordrecht, 1992. 302-340.

Gulosino, Charisse, and Gary Miron. "Growth and performance of fully online and blended K-12 public schools." *education policy analysis archives* 25 (2017): 124.

Haas, Christian, Margeret Hall, and Sandra L. Vlasnik. "Finding optimal mentor-mentee matches: A case study in applied two-sided matching." *Heliyon* 4.6 (2018): e00634.

Halawah, Ibtesam. "The effect of motivation, family environment, and student characteristics on academic achievement." *Journal of instructional psychology* 33.2 (2006): 91-100.

Hannum, Wallace H., et al. "Distance education use in rural schools." Journal of Research in Rural Education (Online) 24.3 (2009): 1.

Hart, Carolyn. The persistence scale for online education: Development of a psychometric tool. Diss. University of Missouri–Kansas City, 2012.

Herdlein, Richard, and Emily Zurner. "Student satisfaction, needs, and learning outcomes: a case study approach at a European university." *Sage Open* 5.2 (2015).

Hodgman, Matthew. (2018). Understanding For-Profit Higher Education in the United States Through History, Criticism, and Public Policy: A Brief Sector Landscape Synopsis. Journal of Educational Issues. 4. 1. 10.5296/jei.v4i2.13302.

Holcomb, Edie L. Getting excited about data: Combining people, passion, and proof to maximize student achievement. Corwin Press, A SAGE Publications Company. 2455 Teller Road, Thousand Oaks, CA 91320, 2004.

Hung, Jui-Long, Yu-Chang Hsu, and Kerry Rice. "Integrating data mining in program evaluation of K-12 online education." *Journal of Educational Technology & Society* 15.3 (2012): 27-41.

Hysenbegasi, Alketa, Steven L. Hass, and Clayton R. Rowland. "The impact

of depression on the academic productivity of university students." *Journal of mental health policy and economics* 8.3 (2005): 145.

Johnson, Monica Kirkpatrick, Robert Crosnoe, and Glen H. Elder Jr. "Students' attachment and academic engagement: The role of race and ethnicity." *Sociology of education* (2001): 318-340.

Jowett, Sandra, and Mary Baginsky. "Parents and education: a survey of their involvement and a discussion of some issues." Educational research 30.1 (1988): 36-45.

Kang, Hyeon-Suk, and Hye-Won Shin. "Reconstruction of Social Science and Humanities Through Narrative." Humanities & Social Sciences Reviews 7.5 (2019): 134-140.

Karcher, M. J. "Meet-n-Greet: A mentor-mentee matching approach for increasing the prevalence of naturally selfselected mentoring partners in program-based matches." *Unpublished manuscript, University of Texas at San Antonio.* Retrieved January 2 (2007): 2008

Kim, Paul, Flora Hisook Kim, and Arafeh Karimi. "Public online charter school students: Choices, perceptions, and traits." *American Educational Research Journal* 49.3 (2012): 521-545.

Kingma, Bruce, and Stacey Keefe. "An analysis of the virtual classroom: Does size matter? Do residencies make a difference? Should you hire that instructional designer?." *Journal of Education for Library and Information Science* (2006): 127-143.

Klein, Stephen, Vi-Nhuan Le, and Laura Hamilton. *Does matching student and teacher racial/ethnic group improve math scores.* RAND CORP SANTA MONICA CA, 2001.

Kominers, Scott Duke, and Tayfun Sönmez. "Matching with slot-specific priorities: Theory." *Theoretical Economics* 11.2 (2016): 683-710.

Landsman, Julie, and Chance Wayne Lewis. *White teachers, diverse classrooms: A guide to building inclusive schools, promoting high expectations, and eliminating racism.* Stylus Publishing, LLC., 2006

Larkin et al, 2015 - Job Satisfaction, Organizational Commitment and Turnover Intention of Online Teachers in the K-12 setting

Lawson, R., and Ann Zerkle. "Price discrimination in college tuition: an empirical case study." Journal of Economics and Finance Education 5.1 (2006): 1-7.

Lee, Kibeom, and Michael C. Ashton. "Psychometric properties of the HEXACO personality inventory." Multivariate behavioral research 39.2 (2004): 329-358.

Lemmon, Lesli Nichole. *Student perception of teacher feedback and the relationship to learner satisfaction in a high school online course.* Diss. Lindenwood University, 2014.

Leslie, Larry L., and Paul T. Brinkman. The Economic Value of Higher Education. American Council on Education/Macmillan Series on Higher Education. Macmillan Publishing, 866 Third Avenue, New York, NY 10022, 1988.

Li, Qing, Lynn Moorman, and Patti Dyjur. "Inquiry-based learning and e-mentoring via videoconference: a study of mathematics and science learning of Canadian rural students." *Educational Technology Research and Development* 58.6 (2010): 729-753.

Liu, Feng, and Cathy Cavanaugh. "High enrollment course success factors in virtual school: Factors influencing student academic achievement." *International Journal on E-learning* 10.4 (2011): 393-418.

Liu, Feng, et al. "The Validation of One Parental Involvement Measurement in Virtual Schooling." *Journal of Interactive Online Learning* 9.2 (2010).

Malave, Eddy R. *The Relationship of Learner-Centered Beliefs of North Carolina Virtual Public School (NCVPS) Teachers and Student Achievement on the North Carolina End-of-Course Assessments.* Gardner-Webb University, 2012.

Marsh, Herbert W., Andrew J. Martin, and Jacqueline HS Cheng. "A multilevel perspective on gender in classroom motivation and climate: Potential benefits of male teachers for boys?." *Journal of Educational Psychology* 100.1 (2008): 78.

Martin, José M., Alvaro Ortigosa, and Rosa M. Carro. "SentBuk: Sentiment analysis for e-learning environments." 2012 International Symposium on Computers

in Education (SIIE). IEEE, 2012.

McCormick, Meghan P., et al. "Estimating Causal Effects of Teacher-Child Relationships on Reading and Math Achievement in a High-Risk Sample: A Multi-Level Propensity Score Matching Approach." *Society for Research on Educational Effectiveness* (2013).

Milner IV, H. Richard. "Race, culture, and researcher positionality: Working through dangers seen, unseen, and unforeseen." *Educational researcher* 36.7 (2007): 388-400.

Morey, Ann I. "Globalization and the emergence of for-profit higher education." Higher education 48.1 (2004): 131-150.

Murray, Joseph, David P. Farrington, and Manuel P. Eisner. "Drawing conclusions about causes from systematic reviews of risk factors: The Cambridge Quality Checklists." *Journal of Experimental Criminology* 5.1 (2009): 1-23.

Mwanza, Alnord LD, George Moyo, and Cosmas Maphosa. "Ethical behaviours of student teachers' mentors in forced same-gender and cross-gender matches in a Malawian Initial Primary Teacher Education programme: implications for mentor selection and development." *Africa Education Review* 14.3-4 (2017): 67-92.

National Centre for Education Statistics, 2020 - `https://nces.ed.gov/fastfacts/display.asp?id=372`

Niche, 2020 - `https://www.niche.com/k12/search/best-online-high-schools/`

Nieto, Sonia. "Placing equity front and center: Some thoughts on transforming teacher education for a new century." *Journal of teacher education* 51.3 (2000): 180-187.

Oliver, Kevin, Shaun Kellogg, and Ruchi Patel. "An investigation into reported differences between online foreign language instruction and other subject areas in a virtual school." *Calico Journal* 29.2 (2012): 269-296.

Ortigosa, Alvaro, José M. Martín, and Rosa M. Carro. "Sentiment analysis in Facebook and its application to e-learning." Computers in human behavior 31 (2014): 527-541.

Packer, Janis, and John D. Bain. "Cognitive style and teacher-student compatibility." *Journal of educational psychology* 70.5 (1978): 864.

Panigrahi, Ritanjali, Praveen Ranjan Srivastava, and Dheeraj Sharma. "Online learning: Adoption, continuance, and learning outcome—A review of literature." *International Journal of Information Management* 43 (2018): 1-14.

Parayil, Govindan. "The digital divide and increasing returns: Contradictions of informational capitalism." The Information Society 21.1 (2005): 41-51.

Rao, Kavita, Michelle Eady, and Patricia Edelen-Smith. "Creating virtual classrooms for rural and remote communities." *Phi Delta Kappan* 92.6 (2011): 22-27.

Ravaglia, Raymond. "Online High School at Stanford University." Understanding Our Gifted 19.4 (2007): 6-9.

Repetto, Jeanne, et al. "Virtual high schools: Improving outcomes for students with disabilities." *Quarterly Review of Distance Education* 11.2 (2010).

Rice, Kerry Lynn. *Priorities in K–12 distance education: A Delphi study examining multiple perspectives on policy, practice, and research.* Boise State University, 2006.

Roberts-Young, Gabrielle Christine. *Does race-matching matter?: An examination of the links between teacher-student racial match and the quality of relationships.* Diss. Rutgers University-Graduate School of Applied and Professional Psychology, 2018.

Rogers, Susan Haley. *Investigating Student Satisfaction and Retention in Online High School Courses.* Diss. 2014.

Shipp, Veronica H. "Factors influencing the career choices of African American collegians: Implications for minority teacher recruitment." *Journal of Negro Education* (1999): 343-351.

Shoaf, Lisa M. "Perceived advantages and disadvantages of an online charter school." *The American Journal of Distance Education* 21.4 (2007): 185-198.

Smith, Lhe. "Differences in Students' Satisfaction of the Economics and

Personal Finance Virtual High School Course Between Students Attending Economically Disadvantaged and Non-Economically Disadvantaged Schools in Virginia ID: 11707." (2016).

Stallings, D. T., et al. "Academic Outcomes for North Carolina Virtual Public School Credit Recovery Students. REL 2017-177." *Regional Educational Laboratory Southeast* (2016).

Syed, Moin, et al. "Individual differences in preferences for matched-ethnic mentors among high-achieving ethnically diverse adolescents in STEM." *Child development* 83.3 (2012): 896-910.

Thompson, Lindsay A., Rick Ferdig, and Erik Black. "Online schools and children with special health and educational needs: Comparison with performance in traditional schools." *Journal of medical Internet research* 14.3 (2012): e62.

Thorne, Kaye. Blended learning: how to integrate online & traditional learning. Kogan Page Publishers, 2003.

Tyndall, Vivian W. *Comparison study: Virtual and traditional classrooms on high school students' mathematics and English academic achievement.* Diss. South Carolina State University, 2014.

van Noort, A. A. A. "Matching teacher feedback and student perceptions in a collaborative learning environment."

Van Vught, Frans. "Mission diversity and reputation in higher education." Higher Education Policy 21.2 (2008): 151-174.

Walker, Christopher O., Barbara A. Greene, and Robert A. Mansell. "Identification with academics, intrinsic/extrinsic motivation, and self-efficacy as predictors of cognitive engagement." Learning and individual differences 16.1 (2006): 1-12.

Wang, Yinying, and Janet R. Decker. "Can virtual schools thrive in the real world?." *TechTrends* 58.6 (2014): 57-62.

Watson et. al. 2014 - `https://eric.ed.gov/?id=ED558147`

White, Doug. Education and the State: Federal Involvement in Educational

Policy Development. Policy Development and Analysis Series. Publication Sales, Deakin University Press, Deakin University, Geelong, Victoria 3217, Australia, 1987.

Whittaker-Coleman, Tanya Rene. *The Effects of Response to Intervention (RTI) on Student Achievement in a Virtual High School.* Trevecca Nazarene University, 2017.

William Blair Equity Research, August 12th 2019, K12 initiation.

Wolfinger, Suzanne. *An exploratory case study of middle school student academic achievement in a fully online virtual school.* Drexel University, 2016.

Young-Jones, Adena, et al. "Bullying affects more than feelings: The long-term implications of victimization on academic motivation in higher education." Social psychology of education 18.1 (2015): 185-200.