Does Upward Bound Have an Effect on Student Educational Outcomes? A Reanalysis of the Horizons Randomized Controlled Trial Study

By

Alan B. Nathan

A dissertation submitted in partial fulfillment of

the requirements for the degree of

Doctor of Philosophy

(Educational Leadership and Policy Analysis)

At the

UNIVERSITY OF WISCONSIN - MADISON

2013

Date of final oral examination: 02/08/2013

This dissertation is approved by the following members of the Final Oral Committee:

Dr. Doug Harris, Associate Professor, Economics (Chair)
Dr. Geoffrey Borman, Professor, Educational Leadership & Policy Analysis
Dr. Felix Elwert, Associate Professor, Sociology
Dr. Adam Gamoran, Professor, Sociology
Dr. Sara Goldrick-Rab, Associate Professor, Sociology

Acknowledgment

It is with great pleasure that I thank all of those who have supported my doctoral research. I would like to thank my advisor Dr. Doug Harris for his eternal patience and practical wisdom. My sincere gratitude goes to the members of my dissertation committee. I was extremely fortunate to have had their encouragement and assistance. Dr. Geoffrey Borman introduced me to randomized control trials and helped me to recognize its strengths and weaknesses. Dr. Felix Elwert guided me through the forest of causal inference and made sure I came out on the other side. Dr. Adam Gamoran motivated me to think about how to integrate theory and practice. Dr. Sara Goldrick-Rab pushed me to understand the connections between policy and intervention.

I would also like to thank the Department of Educational Leadership & Policy Analysis at UW-Madison for their consistent and unyielding financial, academic and social support. A special thank you to Shari Smith, who saw my potential and allowed me to realize it.

Finally I am grateful for the love and affection from family and friends. My dad and step-mom cheered me on as I moved from student, to dissertator, to doctorate. My brother demystified the workings of a graduate school of education as only a member of the inner circle could. My dear girlfriend, and now fiancée became my partner, editor, sounding board, and closest friend. Thank you one and all

TABLE OF CONTENTS

Abstract

The stated goal of Upward Bound (UB) is to increase the rate at which traditionally disadvantaged students graduate from high school and enroll in and graduate from postsecondary educational institutions. Past experimental studies of Upward Bound (e.g., the Horizons study) have found it to be generally ineffective in improving these student educational outcomes. However, my review of the studies revealed apparent methodological and analysis problems. Threats to internal validity are: 1) the exclusion of valid outcome data for between 13% and 19% of students depending upon the outcome measure, and 2) variable imbalances between the treatment and control groups with biases that favored the control group, coupled with the fact that covariate adjustments are not used for all educational outcomes. Problems that primarily threaten external validity are: 1) the use of a sample selection process, which appeared to circumvent a number of eligibility screening criteria normally employed by UB sites, and therefore created a sample that is different from the typical set of program eligible students and, 2) The use of extreme, and possibly inaccurate sampling weights, which greatly increased the variance of the point estimates. Analyses to date have not fully addressed these issues and therefore led to the production of potentially inaccurate estimates. A previous re-analysis of the Horizons study data conducted by the Council for Opportunity in Education (COE), which found treatment effects for post-secondary enrollment and Bachelor of Arts completion rates, suggested similar internal and external validity issues and addressed those issues by removing or reweighting an outlier site prior to estimating the effects of the intervention. However, those solutions drop otherwise valid student responses, use a sample design that is imbalanced and non-representative, and still utilize an arguably incorrect probability weighting scheme. My

proposed solutions to the methodological problems include using all waves of survey data so that fewer observations are dropped, using covariate-adjusted models to account for imbalances, estimating separate effects for students who would be typically eligible for UB, and trimming weights to reduce the mean squared error. With these methods, my intent to treat (ITT) estimates suggests that UB improves the high school and postsecondary outcomes for low-income, first generation students who took part in the experiment. Specifically I found evidence that suggests UB increases high school graduation rates by 4.5 percentage points, post-secondary education enrollment rates by 2.9 percentage points and post-secondary completion rates by 4.6 percentage points. In addition I found some evidence of effect heterogeneity when comparing the treatment and control groups: students who might be typically declared ineligible for UB participation had post-secondary completion rates 8.4 percentage points higher than typically eligible students; the effects are positive but smaller for typically eligible students. In contrast, Horizons study researchers found no evidence of treatment effects on high school graduation or post-secondary enrollment, and did not explore effect heterogeneity according to eligibility. One possible reason for the differences in post-secondary results is that I do not have access to all the post-secondary data sources that were used in the Horizons study. While the results from prior analyses are not robust to all weighting methods, the covariate-adjusted models are more robust. There are two major implications of my findings. First, UB can be used to reduce high school dropout rates. Second, UB eligibility screening processes, such as those that were in place during the time of the Horizons study, should be amended to facilitate the participation of typically ineligible students. Overall, there is compelling evidence that UB can narrow attainment gaps between students from low and high-income households.

List of Tables

Program Definitions

"Educational Talent Search" (ETS) is a TRIO-funded college access initiative that works with partner schools to prepare students for post-secondary education enrollment. Student services include post-secondary education planning, financial aid information, college visits, scholarship data as well as career counseling. ETS aids those aged 11 to 27 from families with incomes under $24,000 (as of 2011) and are potential first time college goers, to gain awareness of educational opportunities.

"GEAR-UP" stands for Gaining Early Awareness and Readiness for Undergraduate Programs. This school-wide, cohort based TRIO initiative was created to increase the college readiness of low-income students who attend high-poverty schools. GEAR UP is a grant-based program that funds the development and delivery of localized services to students in grades 7 through 12. GEAR UP funds are also utilized to fund college scholarships to these students.

"TRIO" is the name given to the original three federal programs (Upward Bound, Educational Talent Search and Student Support Services) created to increase college enrollment and degree attainment for low-income students. Now there are eight TRIO programs (The other five are: The Educational Opportunity Centers program, The Ronald E. McNair Post-baccalaureate Achievement program, Veterans Upward Bound, Upward Bound Math and Science, and The Training Program for Federal TRIO Programs Staff).

"Upward Bound" is a federally funded, after school TRIO program that provides academic support and cultural enrichment to high school students from low-income households and high school students who are first-generation college. The program goals are to increase the rates of high school graduation, post-secondary enrollment and post-secondary completion for the target population.

"Upward Bound Math and Science" (UBMS) is a federally funded TRIO program similar to Upward Bound in many respects including student eligibility profile, average funding per student and the inclusion of summer learning and cultural enrichment sessions. The key points of differentiation are that the instructional components of the intervention place considerably greater emphases on hands-on laboratory science, scientific research methods, and mathematics through at least pre-calculus.

Technical Terms

"Always-takers" are those experiment subjects who seek treatment irrespective of their treatment assignment condition. Control subjects in this category are referred to as "crossovers"

"Average treatment effect" or ATE is the observed mean difference between treatment and control group outcomes. For experiments, ATE and average causal effect (ACE) are the same.

"Balance" is the state where distribution equalization has been achieved and no covariate adjustments are needed to estimate causal effects of treatment (Iacus, King and Porro, 2011). Thus, better matching improves balance.

"Compliers" are those experiment subjects who would receive the treatment if they were randomized into the treatment group, but would not receive the treatment if they were randomized into the control group (Bloom, 2005).

"Complier average causal effect" or CACE is the treatment effect that arises when the subjects in the experiment act in accordance with their treatment assignment. When treatment receipt perfectly aligns with treatment assignment, CACE is the same as ACE and ATE.

"Defiers" are those experiment subjects who act in opposition to their treatment assignment. Defiant control subjects would seek out treatment only if they were denied it and defiant treatment subjects would reject treatment only if they were assigned to it.

"Intent to treat" or ITT is a term used to identify the sample offered the opportunity to receive a treatment. In the Horizons study these are the students who were assigned to Upward Bound irrespective of whether they received treatment.

"Matching" is broadly defined as any method that aims to equate the distribution of covariates in the treatment and control groups (Stuart, 2010).

"Never-takers" are those experiment subjects who reject treatment irrespective of their treatment assignment condition.

"Treatment on the treated" or TOT defines the effect of the intervention on those who received at least the minimum level of treatment. In this study, that level is set at attending at least one Upward Bound session.

Chapter 1 – Introduction

High school graduation, college enrollment and college completion rates for disadvantaged students trail other groups (e.g., Asian, White, not impoverished) by a wide margin. Heckman and LaFontaine (2007) found 65% of Black and Hispanic students graduated from high school as compared to 82% of White students. Other authors find nearly identical results (Stillwell, Sable, and Plotts, 2011). Similarly, the 1996 National Assessment of Educational Progress (NAEP) results for 12th graders shows a Black/White reading score gap of 0.8 standard deviations, and a mathematics score gap of 0.9 standard deviations (Jencks and Phillips, 1998). Although these gaps are shrinking, the differentials remain stubbornly large (Magnuson and Waldfogel, 2011). For example, the Black/White gap for 12th grade reading scores on the NAEP assessment narrowed from roughly 1.3 standard deviations in 1970 to about 0.8 standard deviations by 1996. Similarly, the Black/White gap for 12th grade mathematics scores on the NAEP assessment narrowed from roughly 1.4 standard deviations in 1974 to about 0.9 standard deviations by 1996 (Jencks and Phillips, 1998).

Differences in college enrollment and completion rates across racial groups are also evident, albeit less pronounced. Among 18-24 year olds, approximately 38% of Blacks and 32% of Hispanics enter college as compared to approximately 43% of Whites (Fry, 2010). Among college entrants, the six-year Bachelor of Arts (B.A.) attainment percentages for students who entered college in 2002 was lowest among African–Americans (40%), trailing those for Hispanics (49%) and Whites (60%) (Aud, Hussar, Kena, Bianco, Frohlich, Kemp and Tahan, 2011).

If one recasts the attainment gaps in terms of differing levels of family income, the differences are even larger (Reardon, 2011). Bailey and Dynarski (2011) reported that 80% of all students who are in the highest income quartile attend college while 29% of all students in the lowest income quartile do so, resulting in a college entrance gap of 51 percentage points. Furthermore, they reported that 54% of all students (i.e., unconditioned upon prior college entrance) who are in the highest income quartile complete college while 9% of all students in the lowest income quartile do so, resulting in a college completion gap of 45 percentage points.

Putting these observations together, I see that for every 100 children from high-income families, 80 attend college and 54 complete it, while for every 100 children from low-income families, 29 attend college and 9 complete it. Individuals from the highest income families are roughly 2.8 times more likely to enroll in college and 6.0 times more likely to complete it as compared to persons from the lowest income families.

Higher levels of education are associated with higher lifetime earnings (Mincer, 1974; Cameron and Heckman, 1993; Kane and Rouse, 1995). Therefore, any limits on the education received by disadvantaged groups may have negative economic consequences (Becker, 1962; Mincer, 1974: Cameron and Heckman, 1993) as well as negative health and social outcomes (Wolfe and Haveman, 2002) for these groups.

For much of the last fifty years policy-makers, instructional leaders, researchers and educators have devoted considerable resources to shrinking these educational gaps (Jencks and Phillips, 1998; Harris and Herrington, 2006; Gamoran, 2007; Magnuson and Waldfogel, 2011). One collection of initiatives intended to reduce the size of the problem are pre-college programs. In general, these programs are designed to lessen gaps in educational outcomes by providing low

income and minority students with the academic background and cultural knowledge required for academic success.

Overview of Pre-College Programs

Decades have passed since the launch of the first pre-college pilot programs. These initiatives have now matured to the point where several comprehensive studies were commissioned to catalog the hundreds of studies and describe the landscape of pre-college interventions (Swail, 2001; Gandara and Bial, 2001; James, Jurich and Estes, 2001). These reviews have identified and chronicled common program goals, target populations, key success factors, and challenges found across the various treatments that comprise the pre-college program universe.

These reviews showed that virtually all programs with a post-secondary enrollment or completion goal targeted disadvantaged students, meaning those with one or more of the following characteristics: low-income, minority, or first in their family to attend college ("first generation"). College enrollment and completion were universally listed as program outcomes; although no distinctions were apparently made about the type of college attended by the student (Gandara and Bial, 2001; Swail, 2001; James, et al., 2001). The listed pre-college programs often catered to high achieving disadvantaged students, and only a very small number of programs focused their efforts on struggling students (Gandara and Bial, 2001; Swail, 2001; James, et al., 2001).

Identifying which programs were successful and the source of these successes were complicated by lack of valid comparison groups and paucity of outcome measure data (Gandara and Bial, 2001; Gullat and Jan, 2003; Swail, 2001). Although 94% of programs reported

conducting some type of evaluation, Swail (2001) reported that only 24% of programs even attempted to track college completion.

James et al. (2001) noted that out of a universe of 200 programs aimed at raising minority student achievement only 38 programs collected academic outcome measures. Just 19 out of those 38 studies used a comparison or control group and just four compiled longitudinal data. Arguably these four studies were the most rigorously designed investigations of these interventions and three of the four interventions were either preschool (The High/Scope Perry Preschool Program) or K-12 directed programs that measured the effects of small class size initiatives (e.g., SAGE and STAR). The only post-secondary initiative out of the four studied that employed a randomized controlled trial design and collected academic outcome data was Upward Bound.

Gandara and Bial (2001), James, et al. (2001), and Swail (2001) were each able to find a number of factors that were common to those programs deemed successful by the directors of those same programs. Common success factors included 1:1 or small group instruction or events, a program of study (suggested or delivered) that was academically challenging and was multi-year in duration, and, the inclusion of cultural and pragmatic events that focused on how a student gets accepted to a post-secondary institution (Gandara and Bial, 2001; James, et al, 2001; Swail, 2001).

Swail (2001) among others also noted a number of widespread challenges that might have limited the ability of program administrators to evaluate the success of programs. Limitations on program assessment included high student attrition rates, limited program

duration and intensity, poor or no evaluation metrics and tracking capabilities, and lack of cost data.

A more recent evaluation of college outreach programs found that interventions designed for first generation or low-income students did not appear to increase the educational outcomes of those students (Domina, 2009). Domina (2009) created treatment and comparison groups using propensity score matching, and then contrasted the high school performance and post-secondary trajectories of the two groups to estimate the effects of participation in targeted (i.e. Upward Bound) and school wide (i.e. Talent Search) college outreach program. One limitation of this analysis is that selection effects on the part of program administrators could not be addressed and effect estimates might be understated (Domina, 2009).

Ideally then, an evaluation of precollege programs should eliminate selection effects and measure all major outcomes. Applying these criteria supports the drawing of causal inferences while reducing or removing confoundedness.

Classifying and Appraising Upward Bound

The authors of the earlier comprehensive studies were unanimous in classifying Upward Bound (UB) as an exemplary pre-college program (Gandara and Bial, 2001; Gullat and Jan, 2003; James, et al, 2001; Swail, 2001). In addition to incorporating services that delivered all of the success factors listed above, it was one of the very few programs that had an explicit evaluation component, and it was the only pre-college program to be evaluated using a randomized controlled trial (RCT).

UB is a longstanding federally funded, after-school, pre-college initiative. It is a student-centered program that targets students who are between 13 and 19 years of age, have completed

8th grade, and requires academic support to pursue a postsecondary education. All students must be either low-income or first-generation to college. The initiative is also an event cohort program, meaning that the same UB class can have students from different school grades. UB is a flagship TRIO[1] program, formally legislated as part of the larger Economic Opportunity Act of 1964. Although the program was initially housed within the Office of Economic Opportunity, in 1969 the Office of Education (USOE) was deeded responsibility for the program. Starting in the early 1970's UB's funding authorization came from the amended Higher Education Act of 1965 (Burkheimer, Levinsohn, Koo, and French, 1979).

UB shares a number of limitations common to many pre-college programs. Approximately 37% of students exit the program after less than one year and about 65% of students leave the program prior to high school graduation (Myers and Schirm, 1999). Survey attrition rates are also high, making it difficult to fully ascertain the impact of UB on the post-secondary education enrollment and completion rates of participants. Program costs were reported as an average per student per year, and metrics such as cost-effectiveness were not included (Myers and Schirm, 1999).

Mathematica Policy Research Inc. (MPR) conducted a decade long series of evaluations of UB, collectively known formally as the "National Evaluation of Upward Bound" and informally as the Horizons study, the term I use going forward. The Horizons study evaluations carry considerable weight because these assessments were conducted using RCT. In a properly executed RCT, differences in outcome measures between treatment and control groups are

[1] "TRIO" is the name given to the original three federal programs (Upward Bound, Educational Talent Search and Student Support Services) created to propagate college enrollment and degree attainment for low-income students. Now there are eight TRIO programs, and the name is conventional rather than descriptive (like "The Big Ten").

attributable to the treatment. Studies of this kind are considered to have high internal validity (Shadish, Cook, and Campbell, 2002).

MPR researchers found no evidence that the program has been effective in closing achievement or attainment gaps (Myers and Schirm 1997; Myers and Schirm, 1999: Myers, Olsen, Seftor, Young, and Tuttle 2004; Seftor, Mamun, and Schirm 2009). Specifically, Myers, et al. (2004) found the following regarding high school outcomes:

> For the average eligible applicant, Upward Bound had a statistically insignificant effect on most high school academic outcomes, including total credits earned in the five core subjects— math, science, English, social studies and foreign language—credits earned in honors and Advanced Placement courses, grade point average and high school completion. (p.24)

While Seftor, et al. (2009) found the following regarding post-secondary enrollment:

> Upward Bound did not have a detectable effect on enrollment in postsecondary institutions within approximately seven to nine years after scheduled high school graduation. Approximately 81 percent of treatment group members and 79 percent of control group members attended some type of postsecondary institution (four-year, two-year, or vocational). (p.43)

In addition Seftor, et al. (2009) found the following regarding post-secondary completion:

> Upward Bound had no detectable effect on the likelihood of completing a postsecondary credential in the seven to nine years after high school (effect size = 5 percent). The program did increase the percentage of sample members whose highest credential was a certificate or license, from four to nine percent (effect size = 23 percent).[2] (p.44)

[2] The increase in the percentages of certificates or licenses was not statistically significant at the .05 levels. These data are displayed in Table III.1, page 41 of the report (Seftor, et al., 2009). An alternate way to describe these findings is that UB was estimated to increases post-secondary completion rates by 2.3 percentages points (not significant) and increase certificates or licenses by 4.5 percentages points significant at p=.05).

One consequence of these "null effect" findings has been to threaten the reauthorization of the entire program. For instance the 2005 and 2006 Presidential budgets called for "zero funding" for UB (Field, 2007; Cahalan, 2009: COE, 2012). Although the Horizons study was initiated over twenty years ago, its findings are apparently still relevant.

My findings are different than those produced by the Horizons study researchers. I found that UB had positive causal effects on the educational outcomes of disadvantaged children. My review of the Horizons study revealed significant design and analysis limitations. Threats to internal validity are: 1) the exclusion of valid outcome data for between 13% and 19% of students depending upon the outcome measure, and 2) variable imbalances between the treatment and control groups with biases that favored the control group, coupled with the fact that covariate adjustments were not used in estimating high school outcomes. Problems that primarily threaten external validity are: 1) the use of a sample selection process, which inadvertently circumvented a number of eligibility screening criteria normally employed by UB sites, and therefore created a sample that is different from the typical set of program eligible students and, 2) the use of extreme, and possibly inaccurate sampling weights, which greatly increases the variance of the point estimates. Analyses to date have not fully addressed these issues and therefore might have led to the production of potentially inaccurate estimates. In sum, the methodological choices made by prior researchers significantly influenced the results. I argue that equally reasonable methods yield substantively different results.

A previous re-analysis of the Horizons study data conducted by the Council for Opportunity in Education (COE), which found treatment effects for post-secondary enrollment and Bachelor of Arts completion rates, suggested similar internal and external validity issues and

addressed those issues by removing or reweighting an outlier site prior to estimating the effects of the intervention (Cahalan, 2009). However, COE's approach drops apparently valid student responses, has imbalances between treatment and control, uses a sample design that is non-representative, and utilizes an arguably incorrect probability weighting scheme.

The opportunity to reexamine the original Horizons data has only recently been made available. According to a former Department of Education (USDOE) official, I am the first person outside of the USDOE, MPR or COE to have been given access to the data. This investigation may provide better evidence about how effective UB was in aiding students who participated in the Horizons study reach their academic goals.

Study Significance

Research Contributions

My dissertation provides improved evidence about the effects of UB on high school graduation rates and GPA's, post-secondary education enrollment and completion at select UB sites. My research contributes to the literature on establishing the causal relationship between UB and the educational outcomes of disadvantaged students, and suggests that UB may be effective in narrowing achievement and attainment gaps.

In developing this evidence I identified problems caused by excluding valid outcome data and the sample selection process that have not been previously raised. I also addressed problems previously raised by COE concerning covariate imbalances and weighting schemes, but I solved these problems differently.

Policy Implications

Potential implications for education policy are significant. If UB has no positive effect on educational outcomes then perhaps the money should be spent elsewhere. However, if the intervention causes more students to graduate from high school, and these students are applying to and graduating from post-secondary institutions at higher rates as a result of UB, then perhaps the program should be continued.

Although program funding has overcome past credible threats to its reauthorization, it is still vulnerable to funding cuts based partly on prior null findings. Given the results of this reanalysis, a reassessment of the educational value of UB and a recalibration of the population targeted to receive the intervention may be warranted.

Problem Statement

This dissertation examined the causal connection between UB and student educational outcomes using a longitudinal sample of disadvantaged US high school students who applied to 67 oversubscribed UB project sites during the early 1990's. Participants were studied at three time intervals corresponding to a single pretest period (1992-1994) and two post-test periods (1996-1999 for high school outcomes and 2001-2007 for post-secondary outcomes).

Prior evaluations of the Horizons study data found no evidence of a relationship between UB and the educational outcomes of disadvantaged students. Prior evaluations of the Horizons study data by COE found evidence of a treatment effect on post-secondary enrollment rates and only Bachelor of Arts completion rates. However COE's approach might have introduced another set of problems.

My reviews suggested that those previous evaluations overlooked a number of important and correctable design and analysis flaws. Since UB represents the flagship federal pre-college

program it is important to know if the results of the Horizons study would be different once the flaws were corrected. The findings that emanated from the Horizons study are still relevant and served to influence recent UB funding decisions (Field, 2007; Cahalan, 2009: COE, 2012).

Research Question

Motivated by a desire to address specific criticisms with prior studies and improve the accuracy of the effect estimates, I reanalyzed the newly released data to answer the following questions: Does treatment assignment have an effect on student education outcomes? Does receipt of the treatment have an effect on student education outcomes? Can the findings from the experiment be generalized to the universe of UB students and projects?

In addition to these overarching questions designed to quantify the effect of treatment for those subjects assigned to or receiving treatment, I asked the following related questions: How sensitive are the findings to missing data? Was there evidence of treatment effect heterogeneity across identifiable student sub-groups?

Scope

This dissertation examined the academic outcomes of students who applied to oversubscribed UB sites in the early-1990s. Outcome measures were collected through 2007.

Legal authority to conduct evaluations of TRIO programs was granted under The Higher Education Opportunity Act.[3] Specific types of evaluations now prohibited are those that would require project directors recruit students in numbers beyond what they would normally recruit, or evaluations that would result in the denial of service to an otherwise federally eligible applicant

[3] (HEOA- HR4137), Section 402H, subsection (20 U.S.C. 1070a–18)

(Cahalan, 2009).[4] Put another way, no new artificially induced RCT evaluations of UB like the Horizons study may be legally conducted, but the existing data sets may still be used and are the basis for this re-analysis.

[4] Appendix A, pages 53 and 54

Chapter 2 – Literature Review

Intervention Origins

The genesis of UB can be traced back to philanthropic efforts initiated by the Carnegie, Rockefeller and Ford foundations during the early 1960's. Each foundation was involved in developing pilot programs designed to address the shortage of educational opportunities available to poor and minority students. The Carnegie Foundation's ("Carnegie") approach focused on solving two problems identified through discussions with the American Council on Education: the inadequacy of teacher training given to faculty from Historically Black Colleges and Universities (HBCU), since at that time most minority students attended and graduated from those institutions; and the poor K-12 academic preparation received by students who enrolled at HBCU's. To attend to the latter problem, Carnegie worked in concert with the newly created Educational Service Incorporated to develop pre-college curricula designed to remediate academic deficiencies, particularly those in mathematics, English and science (Groutt, 2011).

The Rockefeller Foundation's ("Rockefeller") plan followed a similarly pragmatic path. Rockefeller sponsored almost two-dozen different pilot programs, resolving to fund those initiatives and program components (e.g., an on-campus summer learning session) that appeared to be effective in providing remedial and enhanced educational opportunities to poor and minority students. In 1963, Rockefeller awarded grants to three institutions, Oberlin College, Princeton University, and Dartmouth College. The pre-college programs offered by those three schools were similar in structure and scope to the UB pilot programs, which followed shortly thereafter. For instance, the grantees targeted their intervention to poor and minority students (both those identified as "talented" and those considered generally disadvantaged), the multi-year course of study offered included academic training and cultural enrichment opportunities starting

as early as the seventh grade, and the curriculum provided students with an intensive, summer-long academic training program (Rockefeller Foundation Annual Report, 1963; Groutt, 2011).

The Ford Foundation's ("Ford") strategy was motivated by the theory of action espoused by Cloward and Ohlin (Groutt, 2011). Cloward and Ohlin thought that persons who are denied access to legitimate means of achieving socially desirable goals might instead use alternative, possibly illegal methods to achieve those same goals. For example an individual attempts to acquire a certain level of education in order to secure a well-paying job however, he is blocked from pursuing that path. So instead he joins a criminal organization that provides education (i.e., "street smarts") and a good source of income. Cloward and Ohlin theorized that if groups of individuals who sought out illegitimate opportunity structures as a means to achieve an end, could instead be redirected into legitimate opportunity structures by the time they entered early adolescence, incidence of crime and other anti-social behaviors among these youths would diminish (Cloward and Ohlin, 1960; Groutt, 2011).

Ford decided to operationalize the tenets of what later became known as Cloward and Ohlin's Theory of Differential Opportunity Systems by offering grants to several institutions that promised to provide summer learning and cultural opportunities for traditionally disadvantaged students. This same line of reasoning also permeated the thinking at the newly developed President's Commission on Juvenile Delinquency, and would later influence the thought process of the Office of Economic Opportunity (OEO), the first home of UB (Groutt and Hill, 2001; Groutt, 2011).

While each foundation took a different route towards developing an operational solution to the problem of differential educational opportunities, the product of their combined efforts helped to frame the content and emphasis of the early UB initiatives. This early literature provides some insight into the genesis of UB. A more general understanding of the relationships between intervention and outcomes requires an explanation of the guiding theories.

Theory of Action and Conceptual Framework

The theory of action for UB is supported by two major sociological theories: the theory of differential opportunity systems (TDOS), and social capital theory. These conceptual frameworks have historically been cited as the theoretical foundations for designing programs aimed at disadvantaged students (Groutt and Hill, 2001; Perna, 2002; Gullat and Jan, 2003; Groutt, 2011). The combined elements of TDOS and social capital theory provide a strong theoretical foundation to explain how UB might be effective in improving educational outcomes for traditionally disadvantaged and first generation students. I discuss each of the theories in turn in the following paragraphs.

Cloward and Ohlin developed TDOS in the 1950's. Their work follows earlier influential sociological theories regarding social order and deviance (Cloward and Ohlin, 1960). The Durkheim/Merton strand (anomie/strain theory) describes the societal and self-imposed "pressures" that can lead to deviance, while the Shaw, McKay, and Sutherland strand (the Chicago School) contains original ideas about how existing social structures set the stage for the development and selection of socially aberrant solutions (Merton, 1938; Shaw, and McKay, 1942; Cloward and Ohlin, 1960).

TDOS posited that society is structured to motivate individuals with a desire to achieve higher economic and social status. However, societies also need institutional structures to support the actions of these motivated individuals. If individuals are prevented or limited in their ability to pursue their goals through legitimate structures an individual may turn to illegitimate structures, which they perceive to offer alternate opportunities (Cloward and Ohlin, 1960). TDOS outlines the courses of action an individual might choose to achieve those goals. For example, an individual may want to achieve higher levels of income and so may choose to enroll in an adult education program if available or, become a member of a street gang as an alternative. Thus, TDOS applies to those individuals who are motivated to achieve higher economic and social status but are prevented from pursuing legitimate means to achieve those goals.

TDOS was used to justify the launch of the earliest UB pilot programs (Groutt and Hill, 2001; Groutt, 2011). While the contributions of the research efforts undertaken by Rockefeller and Carnegie aided in defining some of the early UB program components such as a multi-year summer session, opportunities for cultural enrichment, as well as remedial instruction in math, reading and writing, it was Ford's contribution to the research landscape via adoption of Cloward and Ohlin's body of work, that dominated the theoretical and operational environments (Groutt, 2011). Cloward and Ohlin had been recipients of previous research grants from Ford, and in the introduction to their book "Delinquency and Opportunity; A Theory of Delinquent Gangs" they note that the concepts conveyed within it were rooted in research backed by Ford (Cloward and Ohlin, 1960). This prior relationship might have facilitated the transformation of their ideas from conception to operation.

TDOS is distinguished from other sociological theories about access to legitimate and illegitimate means (e.g., cultural-transmission and differential-association theories, which posit that deviant behavior is learned and transmitted using social interactions) by its explicit assumption of differential access to legitimate pathways (Shaw and McKay, 1942; Cloward and Ohlin, 1960). One cornerstone of this hypothesis is that social position shapes access to legitimate means thus reconciling it with the theory of anomie, which suggests that individuals may resort to deviant behavior when they find that they cannot use socially acceptable methods to achieve socially acceptable goals.

The focus of Cloward and Ohlin's inquiry is centered on the emergence of and affiliation with delinquent subcultures. The authors observed that membership in delinquent subcultures in society are concentrated among lower-class urban males entering adolescence. Furthermore, they noted that the "price" of membership in these subcultures is an embrace of behaviors considered to be socially deviant (e.g., criminal behavior, illicit drug use).

In Cloward and Ohlin's view, deviance represents an individual's attempt to solve a problem that cannot be resolved through social conformity. Faced with this problem of adjustment, the individual responds by searching for an alternative system that provides a solution. This system may be nonconforming, delinquent, or neither.

A second cornerstone of the theory is that both legitimate and illegitimate institutions and structures are available to an individual in this context. It is the relative availability of illegitimate structures that influences his choices. In a sense both legitimate and illegitimate structures are in competition for an individual's allegiance, and whichever one presents the superior perceived opportunity, wins.

The original theoretical foundation that motivated UB as well as the UB program components have evolved over time. Although TDOS was introduced over 50 years ago, opportunity theory, as it is called today, is still germane to deviance and delinquency literature. A Google Scholar search conducted on November 9th 2012 showed a total of 3131 citations of "Delinquency and Opportunity; A Theory of Delinquent Gangs". A total of 572 of these citations were from articles published since 2008 and, of those, 384 contained the keyword "education". For instance, TDOS has been referenced in articles on peer status and high school dropout, examining strain in a school context, historical perspectives of African American males as subjects of education policy, similarities and differences in risk factors for violent offending and gang membership, delinquency and the structure of adolescent peer groups, and the relationship between school engagement and delinquency in late childhood and early adolescence (Staff and Kreager, 2008; Lee and Cohen, 2008; Fultz and Brown, 2008; Esbensen, Peterson, Taylor, and Freng, 2009; Hirschfield and Joseph , 2011; Kreager, Rulison, and Moody, 2011). [5]

Contemporary approaches to social capital theory can be traced to three key authors, Bordieu, Coleman, and Putnam (Wall, Ferrazzi, and Schryer, 1998). Bordieu (1986) defined social capital as membership in a group where each member could draw on the capital accumulated by the entire group. Bordieu held a critical view of social capital. Those who could not become members of a group could also not access the capital owned by that group because the existing group members controlled membership in the group (Bordieu, 1986). New members could join their group with out consent.

[5] Opportunity Theory has also been adapted to cover community response to crime (Skogen, 1989; Becker, 2012)

In contrast to Bordieu, Coleman (1988) described social capital as a system of norms, values, and behaviors. He showed that the development of social capital is linked to the development of human capital (Coleman, 1988). In addition, he argued that acquiring social capital allows individuals to achieve certain goals such as high school graduation, and higher achievement test scores that they could not have achieved without it (Coleman, 1988).

Gullat and Jan's social capital framework follows from the Coleman strand. They used social capital theory to describe how UB is conceptualized to improve student outcomes (Gullatt and Jan, 2003). They posit that UB can build social capital by providing essential academic preparation and positive attitudes and beliefs about college that were not otherwise available to the students (Perna, 2002; Gullat and Jan, 2003). Gullatt and Jan's model makes two assumptions. First, it assumes that UB enrollment leads to the attainment of critical intermediate goals such as better academic performance, which in turn leads students to realize longer-term outcomes such as applying to and being accepted at a post-secondary institution. A second assumption is that the UB support system motivated students to raise their levels of achievement (Gullatt and Jan, 2003).

Linking Theory and Intervention

TDOS and social capital theory give credence to the idea that UB could be effective in positively affecting educational outcomes. Cloward and Ohlin (1960) argued that the creation of an alterative, socially acceptable educational pathway would be attractive to students who could not acquire the schooling they wanted from the schools that were available to them. UB functions as this kind of a pathway for students by providing them with the opportunity to

receive academic instruction and cultural enrichment not available in their local schools (Cloward and Ohlin, 1960; Groutt, 2011).

In line with Coleman (1988)'s argument that social capital yields improved educational outcomes, UB provides students with several ways to build social capital including after-school group academic instruction that increases learning opportunities, academic counseling, study groups, and ACT/SAT test-preparation (Perna, 2002; Gullatt and Jan, 2003). Precollege programs can also act to raise student awareness of college, which also contributes to social capital (Perna, 2002). In addition, students who do not perceive themselves as being college eligible can potentially overcome that hurdle by attending the requisite on-campus summer sessions designed to introduce students to college life and the college application process, including applying for financial aid (Gullatt and Jan, 2003). The summer session functions as another way for students to build social capital by offering them the opportunity to acquire the norms, values and behaviors that are more common to students from college-going families.

UB has evolved from its original form, perhaps as a result of the numerous evaluations of its effectiveness. I will explore the evaluations of UB over time and the corresponding programmatic changes in the following sections.

Early Evaluations of Upward Bound

The Office of Educational Opportunity (OEO) commissioned the first descriptive appraisals of UB while the initiative was still in its pilot phase. The educational consulting firm of Greenleigh Associates conducted the initial evaluation at 22 of the approximately 300 UB project sites in 1967-1969. Using data on student high school graduation and college enrollment rates, they found UB to be "an incredible success story" as 70% of UB high school graduates

enrolled in college versus 50% of US high school graduates nationwide during the 1967-1969 timeframe.

Greenleigh Associates noted several factors that they suggested might have led to biased estimates of the benefits of UB. Apparently, many UB participants did not require additional academic support to increase their chances of attending college. Many participants were already the best academic performers at the high schools where UB recruited. Also, approximately 90% of participants were already enrolled in a college preparatory curriculum at their local high school prior to associating with UB (Greenleigh, 1970). Therefore UB appeared to be more beneficial than it was in reality because no adjustments were made for these apparent differences in pre-existing conditions.

Starting in the early 1970's The General Accounting Office (GAO) conducted its own evaluation of UB and concluded that it was not meeting its stated goal of providing targeted students with the skills and motivation necessary to succeed in education beyond high school. However, GAO researchers only visited a total of 15 sites in 9 states during the 1971-73 timeframe and this sample was not nationally representative of the 330 sites in operation at that time. Additionally, inaccurate information systems prevented GAO researchers from accurately describing the post-secondary outcomes of the sampled students. For example, a GAO audit of the OEO UB college enrollment data used in the study found that 30% of the UB participants OEO reported as being in college had in fact dropped out the year before (Staats, 1974).

In 1973 the federal government commissioned the Research Triangle Institute (RTI) to evaluate UB. This study constituted the first comprehensive, nationwide assessment of UB. UB program goals at that time were: (1) increasing the rate of high school completion; (2) increasing

the rate of entry into post secondary education; and (3) generating the skills and motivation necessary for success beyond high school. RTI measured the effect of the intervention in reaching the first two goals (Davis and Kenyon, 1976).

RTI researchers employed a quasi-experimental research design to evaluate UB. Researchers developed a two-stage stratified sample to create the treatment group. The first stage consisted of stratifying the initial sampling frame of 333 sites (serving 51,755 students) by ethnicity, school size, site location, instructional emphasis, and type of hosting institution (2-year or 4-year college). From that frame, 54 nationally representative sites were selected. These 54 sites were chosen using probability sampling, however the specific process was not documented (Burkheimer, et al., 1976).

The second stage involved student selection. For each of the 54 selected sites two feeder high schools that sent at least four students to UB were chosen. From each of these feeder schools three classrooms representing grades 10, 11 and 12 were picked to form the pool of treatment students. All of the treatment students from within those classrooms were then matched with comparison students from the same feeder schools based on student ethnicity, school grade, low-income status, and degree of academic risk. Matching was many-to-one. There were 3710 students in the treatment group and 2340 students in the comparison group. Data were collected through surveys and official student records.

The RTI researchers contrasted the two groups to identify differences in high school persistence and post-secondary entry rates. They did not find evidence that the intervention was correlated with high school persistence. Approximately 98.6% of UB participants were promoted to the next grade or graduated from high school versus 96.8% for comparison group. This

finding reinforced the finding by Greenleigh (1970) that many high school students did not require UB to increase their chances of attending college, and foreshadowed questions about whether the right kind of students are being targeted by UB site directors (Fields, 2007).

UB was correlated with higher levels of post-secondary entrance percentages, however. Approximately 71% of the treatment group entered a post-secondary institution versus 47% for the comparison group. Although RTI matched treatment and comparison groups on observable characteristics, they were not able to account for selection on unobservable ones nor could they rule out baseline treatment and control differences as a reason for the higher post-secondary entrance rates (Burkheimer, et al., 1976; Moore, Fasciano, Jacobson, Myers, and Waldman, 1997). The specter of selection bias suggests one should interpret the RTI study findings with caution.

Recent Evaluations and the Evolution of the Upward Bound Initiative (1976-1995)

Approximately 20 years later, MPR conducted the next two evaluations of UB. The first of these was a retrospective analysis including the results of an extensive grantee survey and a comprehensive feeder school survey, which documented how UB operations and service offerings had changed since the RTI study (Moore et al., 1997). The retrospective report found that the UB intervention, target population, and objectives all changed in notable ways between its 1976 evaluation and 1995. I have summarized the major changes in Table 2.1.

The 1976 RTI report noted that the interventions were focused primarily on remedial instruction with some resources going to curricular support or enrichment. The 1997 grantee survey showed quite a different picture. By 1993, only 3% of the UB projects focused exclusively on remediation, while 24% of the projects offered remedial services in conjunction

with either curricular support or academic enrichment. The remaining 73% of projects lacked a remedial component, which was a shift away from the original program emphasis (Moore et al., 1997; Myers and Schirm, 1999).

The results of the grantee survey suggested that the services offerings at most UB project sites were similar (In this dissertation I use "sites", "projects" and "project sites" to mean the same thing). The similarities in the elements of the service offering might have been influenced by actions taken by Congress, which took effect in 1995. Congress required academic year programs to meet weekly and make student attendance at these sessions compulsory. In addition, Congress required all grantees to offer a set of core academic subjects consisting of composition, literature, and mathematics courses through pre-calculus, lab science and foreign language instruction as part of the intervention (Moore et al., 1997; Harris and Goldrick-Rab, 2010).[6] The recommended minimum six-week summer session became mandatory, as did the summer bridge program between senior year of high school and freshman year of college. A full 90% of programs were in compliance with this provision by 1997 (Moore, et al., 1997). Perhaps as a result of Congressional action approximately 80% of sites offered their services for three or more years of high school, and 70% of sites had an academic year offering that lasted at least seven months.

Also starting in 1995, two-thirds of recruited students needed to be classified as both low-income (formally defined as family income no greater than 150% of the federal poverty level) and first-generation. The remaining one-third must have had one of those two characteristics. Academic risk was no longer listed as an explicit criterion. In addition, from 1995 forward

[6] However, having the same types of services across sites is not the same thing as offering identical forms of the intervention. I will come back to this point later in the dissertation.

students were required to be a US citizen or a permanent resident (Moore et al., 1997). As a point of comparison, the 1976 RTI noted that 75% of participants were required to be low-income students with high college potential who needed academic and/or psychological support. Students classified as "at risk" (no formal definition) could have made up the remaining 25% of the classroom. In fact, however UB instructors classified 51% of participants as "at risk" in the RTI report (Davis and Kenyon, 1976).

In conjunction with all of the programmatic changes put into place between 1976 and 1995, changes were also made to the program's intended outcomes. The goals of the 1990's version of UB were amended to remove "motivation necessary for success beyond high school" and add post-secondary completion (Moore et al., 1997).

The competitive environment in which UB operated became different as well. Moore (1997), citing Wilbur and Lambert (1991) reported that since the mid-1980's approximately 500 non-federally funded pre-college offerings had come into being, a number roughly equivalent to all UB projects funded for 1995. These other pre-college solutions ostensibly targeted the same students who might apply to UB.

Table 2.1. Evolution of the Upward Bound Program

Dimensions	Upward Bound in 1973	Upward Bound in the 1990's
Types of Activities	Remedial instruction	Curricular Support Academic Enrichment Some remedial services Six week summer sessions Weekly academic services Emphasis on core subjects
Outcomes	High school persistence College enrollment Motivation	High school graduation rates College enrollment and completion rates
Target Population	Low income students with high college potential who need academic and/or psychological support (75% of enrollees) "At-risk" students (25%)	Low-income/First generation college (67%) Low-income OR First generation college (33%) Citizen/permanent resident
Other changes		UB Math and Science -1990

Sources: Myers, D., and Schirm, A. (1999). The Impacts of Upward Bound: Final Reports for Phase I of the National Evaluation. Moore, M. T., Fasciano, N. J., Jacobson, J. E., Myers, D., and Waldman, Z. (1997). The National Evaluation of Upward Bound. A 1990's View of Upward Bound: Programs Offered, Students Served, and Operational Issues. Background Reports: The grantee survey report, target school report.

The vast number of changes to the program mechanics, the rise of competing programs, as well the concrete addition of post-secondary completion to the program's intended outcomes served as partial motivation for USDOE to launch a second study, which was also conducted by MPR. To reiterate, a distinguishing feature of the Horizons study is that researchers utilized a RCT design to support measuring the effects of UB against the 1990's goals. I will delve into the specifics of the Horizons study later in this paper.

The rest of the impetus for the second study came from changes to the federal policy framework sanctioning UB. These amendments to the higher education agenda resulted in significant increases in program funding, and an increased focus on accountability. Once

Congress decided that UB was a stable federal program, it moved to reaffirm its commitment to maintain a secure set of long-term grantees by preferentially funding and regularly evaluating those projects. Starting in 1980 Congress adopted a grant scoring system that favored previously funded grantees. Since the adoption of grant scoring, less than 10% of grantees awarded funds in the prior year failed to receive federal funding in the current year.

Additionally, Congress granted new authorizations to reflect the upswing in college enrollments among poor and minority students. During the 1988-95 timeframe the number of federally funded UB programs increased by 49%. In conjunction with programmatic expansions, Congress also loosened the previously tight reins on UB per student funding. During that same 1988-95 time period, per student funding increased by 75% in nominal dollars (Moore et al., 1997).

Congressionally authorized changes in UB program expansion and in per student funding to address college enrollment gaps between children from low and high income households. As a result of these changes, the percentage of low-income high school graduates (i.e., those whose parents are in the bottom 20% of US household income distribution) attending college nearly doubled, increasing from below 30% in 1972 to above 50% by 1993 (Smith, et al., 1995a). As a contrast, during the same timeframe US college attendance increased by about 50%, a rate demonstrably slower than that experienced by students from low-income households. This difference in growth rates helped raise minority participation in college from 15.4% in 1976 to 21.8% in 1993 (Smith, et al., 1995).

The literature I have described so far tracks the development of UB and highlights how changes in the political climate brought about changes in the programmatic offering. It also outlines previous efforts by researchers to gauge program effectiveness.

Upward Bound in the 21st Century

UB as it exists today is widely used. A list of project sites funded for FY2010-11 revealed that there were a total of 953 active projects, serving over 64,000 students, for an annual direct project cost of in excess of $314 million (USDOE, 2010). This dollar amount did not include unremunerated expenses such as free classroom space, volunteered time, or donated computers and texts, which can boost the per student costs by an additional 40% (Harris and Goldrick-Rab, 2010). At a yearly total cost of well over $5000 per enrolled student, per year, these projects are vastly more expensive than other federal initiatives aimed at students with similar backgrounds and challenges, such as GEAR-UP or Educational Talent Search, which have yearly direct costs of less than $500 per student (Albee, 2005; USDOE, 2012). Given the current budget climate, these substantially higher expense levels put UB squarely in the policy cross hairs.

UB can be thought of as having four components: site, target population, intervention, and expected outcomes. UB is a nationwide program, but each UB site functions autonomously, and recruits its students from its feeder high schools using unique behavioral and academic indicators to establish if a student is eligible for their program. These sites are managed by two- or four-year colleges and universities or by community organizations. The overwhelming majority of project sites have been in operation for over a decade (Moore, et al., 1997). The current target population is low-income (family income is equal to or below 150% of federal

poverty line), first-generation students (self-reported) who have completed 8th grade, are between the ages of 13 and 19, and are classified by local program staff as needing academic support in order to pursue a program of post secondary education.

The current form of the intervention is described by the existence of, and mandatory enrollment in, key program components. Enrolled students must pursue an academic program of study in high school, attend UB academic year sessions including after school and occasional Saturday classes plus SAT/ACT exam preparation and tutoring, and attend UB administered 6 to 8 week summer sessions. Supplemental curricular support including tutoring, mentoring and counseling must also be made available. Students may be enrolled in UB for a period of up to four years. The intensity, focus, and frequency of these sessions are subject to local discretion.

The expected outcomes are to increase the rate at which traditionally disadvantaged students graduate from high school, increase post-secondary enrollment rates, and increase the rate at which these students graduate from post-secondary institutions.

UB is more of a funding stream than a tightly specified treatment, and there are wide variations in content across project sites (Moore, et al., 1997). The characterization of UB as a locally tailored intervention is not new. Prior researchers also found that there was no consistent set of features that described UB projects in general or a successful UB project in particular. For instance the authors of the 1976 Research Triangle Institute (RTI) report, which reviewed the available literature on UB and Talent Search, noted that UB was best described as a diffuse set of treatments received by a diverse student population (Davis and Kenyon, 1976).

Treatment variation carries implications for assessing the effect of an intervention. It is difficult to establish a relationship between treatment and outcome when treatment is not defined

uniformly. For example a survey of UB site directors conducted during 1992-1993 pointed out variations in programmatic emphasis. Approximately 29% of directors emphasized academic enrichment, 20% emphasized academic support, 24% emphasized a blend of academic enrichment and support, 3% emphasized remediation and 24% blended remediation with one or more of the other programs of instruction (Moore, et al., 1997).

The academic course offerings have also changed over time. A comparison of the courses offered during the time of the Horizons study and the courses offered in 2000-2001 indicate a shift towards offering a greater number of math and science courses, more advanced math and sciences courses during the school year, a possibly greater emphasis on foreign language courses, less fine arts and less computer oriented courses (table 2.2).

Table 2.2. Changes in the Upward Bound Academic Course Offerings 1992-94 Versus 2000-2001

Percent of programs offering academic coursework, by subject	Horizons Sample 1992-1993		Upward Bound Students 2000–2001	
Coursework	Academic Year	Summer	Academic Year	Summer
Mathematics				
Pre-Algebra	54	75	37	23
Algebra I	63	94	68	67
Algebra II	64	93	70	77
Geometry	63	94	71	77
Trigonometry	N/R	N/R	39	42
Pre-Calculus	N/R	76	49	52
Calculus	N/R	51	31	28
Integrated	N/R	N/R	31	26
Science				
Earth	N/R	56	49	31
Biology	50	86	69	71
Chemistry	N/R	76	68	68
Physics	N/R	60	53	50
Integrated	N/R	N/R	33	30
Any Foreign Language	N/R	N/R	70	89
English				
Composition	75	98	52	50
Literature	70	96	34	36
Reading	53	80	N/R	N/R
Composition and Literature	N/R	N/R	76	87
Performing Arts	N/R	52	N/R	N/R
Art	N/R	52	N/R	N/R
Speech	N/R	54	N/R	N/R
Physical Fitness	N/R	68	N/R	N/R
Applications/software use	N/R	69	N/R	N/R

Table notes: Horizons data limited to those courses offered by a majority of project sites (Moore et al., 1997). Upward Bound 2000-2001 data reflects the percentage of current participants receiving the service (Cahalan, 2004). N/R = not reported.

Experimental Design

A RCT design was chosen for the Horizons study because study samples constructed via random assignment guarantees that, in expectation, students in each group (i.e., treatment or control) have the same observable and unobservable characteristics, except for access to the treatment. Therefore, if the randomization process was properly executed I can attribute differences in outcomes to differences in treatment conditions.

Randomization provides one solution to "the fundamental problem of causal inference" (Holland, 1986; Morgan and Winship, 2007; Stuart, 2010). Rubin (1974) describes the causal effect as a contrast of potential outcomes. Ideally I would observe the same person under simultaneously administered treatment and control conditions. The difference in potential outcomes would then be the causal effect of treatment. The problem is that it is only possible to observe the potential outcome for those persons assigned to treatment or for those persons assigned to control. I cannot compute the potential outcomes under control for those assigned to treatment nor can I compute the potential outcomes under treatment for those assigned to control (Stuart, 2010). In random experiments, because treatment and control groups are balanced in expectation, one can use the observed potential outcomes under treatment to fill in for the potential outcomes of those assigned to control and the observed potential outcomes under control to fill in for the potential outcomes of those assigned to treatment, thus solving the problem.

The research design for the Horizons RCT study was executed in two stages. First, the researchers created a stratified random sample of projects they considered to be nationally

representative of UB projects, and then they randomly selected applicants from those projects into treatment and control groups.

Starting from the universe of 568 operating UB projects, researchers identified a subset of 395 UB projects, which made up the sampling frame. In order to qualify, the project site had to be managed under the direction of a local two or four-year college, with at least three years of hosting experience (Myers and Schirm, 1997). From the cluster of 395 qualifying projects researchers stratified the sample into 46 strata and then randomly selected the 67 projects used in the experiment. One additional criterion imposed upon these 67 projects was that the projects were required to have at least two applicants for each available opening (i.e., they were oversubscribed).

The 46 strata were generated from combinations of four variables. These variables were location of host institution (urban or rural), type of host institution (public four-year, private four-year, two-year), project size (small = 60 or fewer students, medium = 61-99, large = 100 or more), and historical racial composition of the project.[7] MPR researchers created sampling weights to account for the stratified design, which I discuss later in more detail.

Four-year institutions comprised the vast majority of host organizations (Myers, Olsen, Seftor, Young, and Tuttle 2004). Within a given stratum, each project had the same probability of being selected, but selection probabilities across strata were unequal. Small and large sites as well as those administered by two-year colleges were oversampled ostensibly to permit subgroup analyses (Myers and Schirm, 1997).

[7] Historical racial composition was used as a stratifying variable in 63 out of the 67 sites. The number and type of racial composition variables varied from site to site.

The MPR researchers asserted that these 67 oversubscribed UB sites (with stratum weights) were representative of the 395 sites that made up the original sampling frame (Myers and Schirm, 1997; Myers and Schirm, 1999). Oversubscription at the 67 chosen sites may have occurred naturally, or it may have been induced (e.g., by relaxing typical admissions standards) in order to comply with the requirement to have least two applicants for an available opening (Cahalan, 2009). Reacting to the use of the oversubscription criterion, which was considered by most legislators to be tantamount to a denial of services, Congress passed legislation in 2008 (as part of the re-authorization of the Higher Education Opportunity Act (HEOA- HR4137) prohibiting future active over-recruitment at TRIO sites for the purposes of conducting a random assignment evaluation (Cahalan, 2009). This makes analysis of the Horizons Study data particularly important.

Stage two of the random sampling procedure focused on student assignment within the selected sites. Students who applied for acceptance into UB were required to complete background surveys and agree to release their 9th grade transcripts prior to assignment. A number of UB site directors argued against randomization without first blocking on variables such as feeder school or gender. These site directors were allowed to create their own blocking schemes subject to there being "enough" eligible applicants to support randomization within a stratum. Those sites that did not use a blocking scheme used simple random assignment to build their sample. Randomization, both with and without blocks resulted in the creation of 339 random assignment strata,[8] with an average of eight students per stratum (Myers and Schirm, 1999;

[8] Definitions for each of the 339 strata were not published. Thirty of the 339 strata ended up lacking either a control or treatment case (Myers and Schirm, 1999). These strata were collapsed into the nearest neighboring strata as measured by their propensity scores.

Cahalan, 2009). The blocking procedures as well as decisions about the number of students per stratum were not documented and this creates some problems later in trying to address certain limitations of prior analyses.

In addition to the 2844 students who were part of the experiment, another 184 students were non-randomly selected into UB. All 184 students were designated as "must-serve" cases were therefore excluded from the research sample prior to randomization (Myers and Schirm, 1997; Cahalan, 2009). The final sample group was composed of 1524 treatment and 1320 control students (Myers and Schirm, 1999). The treatment group was made larger than the control group because MPR researchers and the site project directors knew there would be some "no-shows" and some early dropouts. Rather than use a wait list, the decision was taken not to refill any open slots and so the treatment group was made larger than the actual number of initial slots and thus larger than the control group.

The size of the sites ranged widely. The largest of the sites had 96 students in the experiment (50 treatment and 46 control), while the smallest had four (two treatment and two control). Mean size of the experimental sample at a given site was 42 students, with a standard deviation of 20 (Myers and Schirm, 1997).

In the Horizons experiment, each student in the sample was weighted so that the sample could represent the population of UB students encapsulated by the sampling frame. First, students were assigned stratification weights. This weight is the inverse of that student's selection probability for the stratified random sample (Myers and Schirm, 1999). Within a given stratum, all students received the same weight (weights could and did vary across strata). For example if the student applied to a project site within the "large, urban, hosted by a 2-year

university" stratum, the selection probability was 1/3 since there were a total of 3 projects in that universe and 1 project was randomly selected into UB. That student was given a weight of 3 to represent that student at the selected project plus the 2 students who applied to project sites not randomly selected to participate in the experiment (Myers and Schirm, 1999).

Next, MPR researchers compared the sample and population distributions of demographic variables including gender, race, student educational aspirations, grade level at program application and federal eligibility for UB (i.e., low income and/or first generation) and weighted the sample observations so that the distributions of the sample demographic variables mimicked the distributions of the population demographic variables. This process gives us a set of post-stratification weights.

In addition researchers adjusted the post-stratification weights described above to reflect survey non-response rates, account for design effects, and support external validity. MPR researchers created one set of non-response weights for high school graduation, and one set for post-secondary enrollment and completion (Myers, Olsen, Seftor, Young, and Tuttle 2004; Seftor, Mamun, and Schirm, 2009).

MPR researchers conducted tests of mean differences (with post-stratification weights) on a number of background variables to determine if application of the two-stage process resulted in the creation of balanced treatment and control groups. The conclusion of the research team was that the randomization process was successful as measured by the results of the means comparison tests. Together with the assumption of no imbalances on other variables, this conclusion implies that the effect estimates should have a high degree of internal validity. The results from a series of t-tests, published as Table B.1 in the first Horizons evaluation report,

show no statistically significant differences at p=. 05 between treatment and control baseline variables for 43 out of 46 variables tested (Myers and Schirm, 1997). [9,10]

These steps constitute the primary design and methodological steps taken by MPR researchers prior to the estimation phase.

Horizons Study Findings and COE Response

MPR researchers tracked the academic progress of the Horizons sample group members for a decade and published four progress reports over the ten-year period. They was unable to find evidence of any impact of the treatment on high school graduation rates, high school GPA, post-secondary enrollment or post-secondary completion (Myers and Moore, 1997; Myers and Schirm, 1999; Myers, et al., 2004; Seftor, Mamun, and Schirm 2009).

The manner in which the Horizons study was conducted generated numerous critical responses dating back to the RCT planning stages in the early 1990's. Objections came from such diverse quarters as the US Congress, the Council for Opportunity in Education (COE), and the Assistant Secretary of Education (Cahalan, 2009; COE, 2012). These objections largely lay dormant until after the publication of the final Horizons study evaluation report, which focused on post-secondary outcomes. After reviewing the experimental design parameters detailed in the final Horizons study report, COE decided to reanalyze the Horizons study data and replicate the published findings.[11]

---

[9] Control group subjects talked more often with their parents about their studies (2.0 versus 1.9) Treatment group subjects played video games more often on the weekend (1.4 versus 1.3), and control group subjects had a slightly higher number of geometry credits (1.1 versus 1.0)

[10] Only one of these 46 variables, Student educational expectations, was used by MPR in their covariate adjusted estimation equations. I used this variable as well.

[11] The Horizons research team defined "impact" as a variable being statistically significant at the .05 levels.

COE was unable to replicate the point estimates and conclusions as published in the final evaluation report even though their researchers had access to the same data files (COE, 2012). COE estimated treatment effects for post-secondary enrollment and completion of a four-year college degree, but did not study high school outcomes (Cahalan, 2009). Although COE did estimate treatment effects for other post-secondary credentials, the main body of the post-secondary completion analysis focused on BA attainment. COE found, in contrast to the Horizons study findings, positive and significant effects of treatment assignment (6.9 percentage points) and treatment receipt (10.9 percentage points) on post-secondary enrollment. Neither UB assignment (0.9 percentage points) nor treatment receipt (1.7 percentage points) had a statistically significant impact on BA attainment, and these findings were consistent with those published in the Horizons study final report.

During the replication process COE found what they thought to be four flaws in the Horizons experiment design. COE thought that one project site (project 69), which comprised 26% of the weighted sample but only 3% of the non-weighted sample should be dropped from the analysis because it was an outlier. COE argued that this site did not recruit or educate students that were representative of the stratum to which project 69 had been assigned. Project 69 was placed in a stratum reserved for urban project sites of medium size that were administered by a 4-year university even though until just prior to the study, the college administering the program was chartered as a two-year college, and many of its major programs were still vocational in nature. Also, the racial composition sub-stratum for this project was designated as "other" meaning no ethnic group comprised a majority of the sample, but in reality the student

sample for project 69 was 58% African-American, 41% Hispanic and 1% "Other" (Cahalan, 2009).

When the COE research team excluded project 69 data from the analysis they produced larger, statistically significant treatment effects.[12] Treatment assignment had a positive impact on for post-secondary enrollment (9.1 percentage points), as did treatment receipt (14.2 percentage points). UB impacts on BA completion were higher (3.7 percentage points for treatment assignment and 7 percentage points for treatment receipt) and statistically significant.

The issue raised by COE concerning the placement of Project 69 may indeed be valid. The U.S. Department of Education's Office of Policy and Program Studies Service posted a series of caveats explicitly cautioning users of the final evaluation report for the Horizons study, that study results were vulnerable to the stratum placement of Project 69 (PPSS, 2009).

COE researchers also noted multiple variable imbalances. Within project 69, control group members had statistically significantly higher mean values on family income and respondent's educational expectations. These imbalances in COE's view could introduce bias since both variables are typically positively associated with academic outcomes (Cahalan, 2009). Imbalances within some sites is not atypical especially when the number of sites is large, therefore the observed imbalances do not constitute a reason for dropping those data.

COE researchers also found that the school year most recently completed by the applicants did not balance between the treatment and control students in the overall sample. Treatment students as per COE's analysis had completed a lower grade on average at time of application by 0.25 academic years (grade at time of application for treatment students was 9.62

[12] MPR claimed that the null effect finding for UB was valid both with and without project 69 (Seftor, et al., 2009).

and grade at time of application for control students was 9.87), which might introduce bias as per COE (Cahalan, 2009). COE's solution to the completed school year problem was to create new outcome variables that were standardized by the expected year in which a student graduated from high school.

Finally, COE argued that the National Student Clearinghouse (NSC) data should not be used as part of the evaluation process as at that time institutional coverage was too low to be reliable, especially for community colleges (Cahalan, 2009).

In sum, there were four methodological issues raised by COE researchers, which they thought obscured the actual effects of UB on educational outcomes. First, they thought that the inclusion of highly weighted project 69 in the sample served to negatively bias the study outcomes. COE observed that that this one project site, which made up 26% of the weighted sample was apparently not representative of the stratum to which it was assigned, and would not have received the same weight if properly situated (Cahalan, 2009). Second, they argued that unaddressed imbalances between the treatment and control groups that favored the control group also exerted a downward bias on educational outcomes. Third they used the Horizons study post-stratification weights (unadjusted for non-response rates) in their estimation equations, which they argued created a more accurate depiction of the national distribution of UB project sites. Fourth, COE argued that NSC data were not usable due to under coverage. The concerns raised by COE have some merit, however, I show through my re-analysis of the data in Chapter 4 that these issues have still not been addressed to the extent possible.

Chapter 3 – Data and Methods

In this chapter I include a description of the sample data, and highlighted differences between the analysis samples as created by MPR researchers and the analysis samples I constructed. I describe the mechanisms by which the data MPR researchers and I used were produced and I develop a series of descriptive statistics that summarizes the data. I detail the threats to internal and external validity and explain how I will mitigate those threats. Next I explain the necessary steps to replicate results from prior studies and produce my effect and impact estimates and test my hypotheses.

Sample Description

I was granted access to the analysis files created for the Horizons study through the U.S. Department of Education's Office of Policy and Program Studies Service (PPSS). These were the files used to generate the Horizon study evaluation reports and the COE report. The sample had been previously de-identified. Data obtained from external sources (i.e., some supplemental Student Financial Aid data and all National Student Clearinghouse data) that were part of the Horizons study and COE analyses were deleted from the released data by PPSS.

The Horizons study sample contained 2844 observations, 1320 control subjects and 1524 treatment subjects from 67 oversubscribed UB project sites (table 3.1). The variables included in the table come from the Horizons study and COE covariate adjustment models, and variables identified in the literature to be associated with educational outcomes (Alexander, Entwisle and Horsey, 1997; Rumberger and Lim, 2008). Participants in the Horizons study were middle and high school students in grades 8-11 who applied for a slot at a designated, oversubscribed UB project site between May 1992 and March 1994. A review of the Horizons data showed that the majority of applicants were; female (67%), minority (62% Black or Hispanic) had applied to UB

in 9th or 10th grade (74%), from low-income households (85%) and had non-college graduate parents (94%). On average, these students expected to graduate from a four-year college (8.2 on a 10-point scale), had taken Algebra or above in the 9th grade (65%) and had a 9th grade GPA of above 2.5 (52%).

Table 3.1. Descriptive Statistics For The Pretest Variables

Variable	Source	N	M	SD	Min	Max
Treatment assignment	Analysis file	2844	0.54	0.50	0	1
Student educational expectations	Baseline survey	2598	8.20	1.76	1	10
Low-income household	Eligibility form	2844	0.85	0.36	0	1
Quarter of birth	Eligibility form	2842	71.85	4.06	41	88
Female	Eligibility form	2844	0.67	0.47	0	1
White	Analysis file	2844	0.28	0.45	0	1
Hispanic	Analysis file	2844	0.19	0.39	0	1
Black	Analysis file	2844	0.43	0.50	0	1
Other race	Analysis file	2844	0.10	0.30	0	1
At least one parent has a B.A.	Eligibility form	2844	0.06	0.23	0	1
Director expects to serve this student	Eligibility form	2835	1.58	0.66	0	3
Student previously received precollege services	Analysis file	2813	0.31	0.46	0	1
Student is attending the same school as last year	Analysis file	2805	0.64	0.48	0	1
Student is living at the same address as last school year	Analysis file	2815	0.83	0.38	0	1
Student applied to UB in 8th grade	Eligibility form	2844	0.15	0.35	0	1
Student applied to UB in 9th grade	Eligibility form	2844	0.39	0.49	0	1
Student applied to UB in 10th grade	Eligibility form	2844	0.35	0.48	0	1
Student applied to UB in 11th grade	Eligibility form	2844	0.11	0.31	0	1

Table notes: Horizons sample data are from student eligibility forms and baseline surveys. Horizons student responses were taken at time of application.

The baseline survey, which served as the repository for most of the pretest data, was at least partially filled out by 99% of applicants (2820 students, 1311 control subjects and 1509 treatment subjects). Only those students who completed the baseline survey and made their high school transcripts available to the research team were supposed to be eligible for randomization and this stated precondition explained the high participation rate. A comparison of the Horizons

sample demographic data from 1992-1993 with available demographic data from a census of 55,140 UB participants from 2000-2001 suggests a demographically stable population (table 3.2).[13]

Table 3.2. Demographic Comparison Between Mean Proportions of Horizons Study Applicants (1992-1993) and Mean Proportions of a Census of UB Students (2000-2001)

Variable	Horizons Sample 1992-1994	Upward Bound Students 2000–2001
Low-income household	0.85	0.84
At least one parent has a B.A.	0.06	0.05
Female	0.67	0.64
White	0.28	0.25
Hispanic	0.19	0.19
Black	0.43	0.45
Other race	0.10	0.11
8th grade	0.15	0.17
9th grade	0.39	0.39
10th grade	0.35	0.33
11th grade	0.11	0.10

Table notes: Horizons sample data are from student eligibility forms and baseline surveys. Horizons student responses were taken at time of application. Upward Bound 2000-2001 data are from Cahalan and Curtin (2004). Upward Bound student responses were taken during the time the students was enrolled in Upward Bound.

Survey Instruments

A consortium that included MPR, ETS, Westat, and Decision Information Resources created the survey instruments used in the Horizons study under contract with the USDOE. I used two instruments, the initial student eligibility form and the baseline survey (in concert with student 9th grade transcript data) as the source for all the pretest measures used in this study. I

[13] I conducted an ad-hoc comparison between Horizons students and students who attended UB feeder schools from 1988 (using NELS: 88) to see if students who volunteered for UB placement might appear to be different than those who were federally eligible (table A.1). Federally eligible students seemed to have lower educational expectations in spite of higher parental educational attainment. Precise comparisons were not possible because NELS: 88 lacked a variable to determine federal income eligibility and because NELS: 88 students were all in 8th grade at baseline while Horizons students were in various grades.

used four instruments, the 2nd, 3rd, 4th and 5th wave student surveys as my data sources for student outcome measures.

A one-time grantee survey was administered to the UB project directors during 1992, which I used as the source for general information about the students who received UB services in the three years preceding the launch of the Horizons study. This grantee survey was administered to the 67 projects used in the Horizons experiment as well as 188 other project sites (Moore, et al., 1997). MPR managed the administration of this survey and produced the descriptive reports.

Study Variables

The pretest study variables used to answer the posed research questions are detailed in Table 3.1. With three exceptions these major independent variables are coded as indicators, where 1=Yes and 0=No. For the "student educational expectations " variable a value of "1" meant the student did not expect to graduate from high school, while a "10" meant the student expected to earn a PhD, MD or other professional degree. I recoded the "Don't Know" responses given in the original survey for this variable to missing, which resulted in 215 missing values.

A categorical variable denotes whether the Project Director would have expected to serve a given student in non-experiment years. Students who in the opinion of the local project director were most likely to be served are coded with "1" and least likely "3". This variable was supposed to act as a rough measure of eligibility for a given UB site, but I found this variable to be uncorrelated with any published eligibility criteria (Moore, et al., 1997).

Finally, I created a continuous "quarter of birth" variable to determine when students were born and to check student age for balance. I coded birth quarters using January 1960 as the

starting point from which to gauge birth dates. So a student born in January 1970 receives a value of 40 while one born in January 1980 receives a value of 80.

Threats to Validity

Survey Responses

The number of responses used in the Horizons study reports is displayed in column 1 of table 3.3. The numbers used in generating my new effect estimates are shown in column 2, and the difference is displayed in the 3rd column. I explain the differences in the following paragraphs.

The MPR researchers used data compiled from the 3rd survey wave (1998-9) when estimating the treatment effects for high school graduation.[14] The consequence of using a single survey wave to establish student educational outcomes is worth noting. If a student reported her high school graduation outcome in the 2nd survey wave, but did not respond to the 3rd survey wave, her response from the 2nd survey wave was ignored (she received a 3rd survey weight of zero).

In contrast to the method used to count high school outcomes, MPR researchers did use all available 12th grade transcript data when estimating the effect of UB on 12th grade GPA. These are the same transcripts that were used in verifying high school graduation outcomes (Myers, et al., 2004). Therefore it is likely that the number of high school outcomes I report in table 3.3 (2645) is more accurate than the number reported by Myers et al. (2291). This difference in the number of high school outcomes also has implications for the accuracy of the

[14] To establish if a student graduated from high school MPR researchers used a combination of high school transcripts obtained directly from the school, telephone interviews, in person and mailed survey responses and reports from the UB site directors (Myers et al., 2004).

non-response weights used in estimating the effects of UB on high school graduation, which I discuss later.

One other assumption I make for this dissertation is that students who reported that they have dropped out of high school in the 2nd survey wave, but did not respond to the 3rd or 4th survey waves (i.e., the other survey waves that asked about high school completion) were considered to have dropped out of high school for good.

In addition, data compiled from the 5th survey wave (2003-4) was used in estimating the treatment effect on post-secondary education outcomes (Myers, et al., 2004; Seftor, et al., 2009; Cahalan, 2009). If a student reported her post-secondary outcome in the 4th survey wave, but did not respond to the 5th survey wave, her response was also ignored (Cahalan, 2009; PPSS, 2011).

The procedure for establishing post-secondary results was complex. Specifically, as Seftor et al., (2009) notes:

> Each analysis is conducted using post-stratification adjusted weights that account for sample selection probabilities and survey nonresponse. In this chapter, we focus on one measure of enrollment (designated 5B in the appendices) and one measure of completion (designated 7B in the appendices). With Measure 5B for enrollment, we code a sample member as an enrollee if he or she is found to be an enrollee in the full NSC data or is a Pell recipient according to the SFA data or said in the survey that he or she was enrolled at some time. The sample member is not an enrollee if he or she does not appear in the NSC data (and is therefore not an enrollee) and has not been a Pell recipient and said in the survey that he or she had never been enrolled. This leaves uncoded the survey nonrespondents who are not in the NSC data and did not receive a Pell grant. For them, we assume that they are not enrollees if they never applied for financial aid. If they did apply for financial aid, we code their enrollment status as missing. The sample members with missing enrollment status get dropped from the analyses of enrollment, and weights for the remaining sample members are adjusted to compensate, as described in Appendix A.
>
> For measuring completion, the SFA data do not provide information on the actual receipt of degrees, certificates, or licenses. Recognizing this limitation in constructing Measure 7B for

completion, we code a sample member as a completer if he or she is a completer according to the full NSC data or said in the survey that he or she has completed a degree, certificate, or license. The sample member is not a completer if he or she has no evidence of completion in the NSC data and said in the survey that he or she had not completed a degree, certificate or license. This leaves uncoded the survey nonrespondents who have no evidence of completion in the NSC data. For them, we assume that they are not completers if they never applied for financial aid. If they did apply for financial aid, we code their completion status as missing. The sample members with missing completion status get dropped from the analyses of completion, and weights for the remaining sample members are adjusted to compensate. (pp. 41-42)

To sum up, MPR appears to prefer the survey data to NSC and SFA data. Post-secondary enrollment (measure 5B) is composed first and foremost of survey responses to the 5th survey wave, supplemented by data from NSC and SFA. Post-secondary completion (measure 7B) is composed primarily of survey responses to the 5th survey wave, supplemented by data from NSC. The SFA data is used to classify a missing survey responder as a non-completer, but cannot be used to classify a missing survey responder as a post-secondary completer.[15]

In contrast, my reanalysis incorporated student responses for survey waves 2-4 when estimating treatment effects for high school graduation and student responses for survey waves 3-5 when estimating treatment effects for post-secondary outcomes.[16] Specifically, for high school outcomes I started with the survey wave 3 and coded students as having graduated, not graduated or were missing based upon their response to a question about if they had yet graduated or not. For students who were coded as missing, I checked survey wave 2 to see if they responded to a question about if they had yet graduated or not. If they responded to that question

[15] It is also quite possible that the SFA data are less valid for describing the two-year sector because students who apply to two-year schools are less likely to apply for financial aid.

[16] Cahalan (2009) made a similar adjustment to the post-secondary outcome data, but did not review high school outcome data.

(and were initially coded as missing) I recoded them. The final step was to review survey wave four to see if I code recode any students I still had coded as missing.

The same logic holds for post-secondary outcomes. Starting with survey wave 5 I coded student responses to questions about enrollment or completion as yes, no or missing. I checked survey wave four for to see if any students I initially coded as missing had indicated their enrollment or completion status. Then I followed the same process for survey wave 3 data.

As a general rule, statistical power is increased by using as much available data as possible, therefore I increased the pool of students who reported their high school outcomes by 354, the pool of students who reported their post-secondary enrollment status by 558, and the pool of students who reported their post-secondary completion outcomes by 793 subjects, relative to the Horizons sample. These changes increased the sample by between 15 and 46 percent above those found when using a single survey wave.

Table 3.3. Differences Between Horizons Study and All Survey Responses

Outcome Measure	Horizons Survey Responses	All Survey Responses	Absolute Difference
High School Graduation	2291	2645	354
12th Grade GPA	2673	2673	0
PSE Enrollment	2102	2660	558
PSE Completion	1724	2517	793

Table notes: Survey response data comes from the Horizons data tables. PSE = post-secondary education. GPA= Grade Point Average. The variables used for high school graduation, PSE Enrollment, and PSE completion responses are TSHGRD3, which captures responses from the 3rd survey wave only, and V5M1A, and V5DEG1A, which capture responses from the 5th survey wave only and exclude SFA and NSC data. My high school graduation, PSE Enrollment, and PSE completion responses are calculated by summing up all unique responses across survey waves 2 through 4 for high school outcomes and 3 to 5 for PSE outcomes and also exclude SFA and NSC data as described in the text.

In most respects, the Horizons method seems like a reasonable plan for imputing missing survey data. Since it counts enrollment on any measure, it might tend to over-state enrollment levels, but there is no reason to expect a differential by control and treatment. However, recall that by using more than one survey wave, the number of observations whose missing values are replaced is quite large. This leads to an internal contradiction in the method: the researchers seem to prefer the survey measures over the SFA and NSC in the 3rd and 5th waves (they only replace missing survey responses), but they seem to prefer SFA and NSC over the other data from other survey waves. Later, I discuss the missing data and replacement rates under various scenarios.

Sample Balance

Differences in baseline characteristics between treatment and control groups might also represent a threat to internal validity. Tables 3.4 through 3.6 depict treatment and control group balance tests using three different weighting assumptions. Variables that are shown to be imbalanced should be accounted for in the estimation equations, especially in cases where the imbalanced variables have been previously show to be associated with the measured outcome. I considered a variable to be imbalanced if the mean difference was significant when using a critical value of .05 or if the standardized mean difference was greater than .25 standard deviations (Ho, et al., 2007).

Table 3.4 shows balance test results absent any weights. I found two variables, student educational expectations and low income, to be out of balance, with imbalances that favor the control group. Put differently, if these imbalances were not adjusted prior to estimation, there would be a downward bias exerted on the treatment effect. All other variables in the sample appear to be well balanced between the treatment and control groups.

Table 3.4. Balance Test Results for the Horizons Study Sample – No Weights

	Control		Treatment		Differences		
Variable	M	SE	M	SE	t	p	SMD [1]
Student Educational Expectations	8.29	0.05	8.14	0.07	2.11	0.03*	0.09
Low Income Household	0.84	0.01	0.86	0.01	1.87	0.06+	0.05
Quarter of Birth	71.89	0.11	71.81	0.15	0.48	0.64	0.02
Female	0.68	0.01	0.66	0.02	1.23	0.22	0.04
White	0.27	0.01	0.28	0.02	0.57	0.58	0.02
Hispanic	0.18	0.01	0.19	0.01	0.20	0.84	0.03
Black %	0.44	0.01	0.42	0.02	1.04	0.30	0.04
Other race	0.10	0.01	0.11	0.01	0.61	0.54	0.03
At least one parent has a B.A. or higher	0.06	0.01	0.05	0.01	0.68	0.50	0.04
Project Director expects to serve this student	1.60	0.02	1.56	0.03	1.61	0.11	0.06
Student previously received Precollege services	0.32	0.01	0.30	0.02	1.58	0.11	0.04
Student is attending the same school as last year	0.65	0.01	0.63	0.02	0.79	0.43	0.04
Student is living at the same address as last school year	0.84	0.01	0.82	0.01	1.44	0.15	0.05
Student applied to UB in 8th grade	0.16	0.01	0.14	0.01	1.22	0.23	0.06
Student applied to UB in 9th grade	0.40	0.01	0.39	0.02	0.39	0.70	0.02
Student applied to UB in 10th grade	0.34	0.01	0.36	0.02	1.20	0.23	0.04
Student applied to UB in 11th grade	0.11	0.01	0.11	0.01	0.14	0.89	0.00

Table notes: Asterisks indicate that the mean differences are statistically significant at p= .05 or lower, a condition to be addressed in the estimation phase. "No Weights" means that each observation receives a weight equal to 1.
[1] Standardized mean differences. Calculated by dividing the absolute difference in group means by the treatment group standard deviation (Stuart, 2010)
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3.5 shows balance test results using the post-stratification weights. These weights are the ones used by COE researchers in their estimation equations of post-secondary impacts. The student educational expectations variable remains imbalanced in this instance. The remaining variables seem to be balanced. These weights do not account for survey non-response in the dependent variables.

Table 3.5. Balance Test Results for the Horizons Study Sample – Using Post-Stratification Weights

	Control		Treatment		Differences		
Variable	M	RSE	M	RSE	t	p	SMD [1]
Student Educational Expectations	8.26	0.10	7.90	0.32	2.37	0.02*	0.21
Low Income Household	0.83	0.02	0.86	0.03	0.20	0.45	0.00
Quarter of Birth	71.80	0.18	71.38	0.02	1.32	0.19	0.11
Female	0.71	0.02	0.68	0.03	1.21	0.23	0.09
White	0.20	0.01	0.21	0.03	0.75	0.45	0.05
Hispanic	0.23	0.02	0.22	0.01	0.42	0.67	0.02
Black %	0.51	0.02	0.50	0.01	0.20	0.85	0.00
Other race	0.06	0.01	0.07	0.01	0.47	0.64	0.04
At least one parent has a B.A. or higher	0.05	0.01	0.05	0.01	0.56	0.57	0.00
Project Director expects to serve this student	1.61	0.03	1.61	0.04	0.00	0.99	0.00
Student previously received Precollege services	0.30	0.02	0.29	0.03	0.14	0.88	0.02
Student is attending the same school as last year	0.59	0.02	0.63	0.62	0.85	0.40	0.06
Student is living at the same address as last school year	0.82	0.02	0.82	0.82	0.00	0.96	0.00
Student applied to UB in 8th grade	0.13	0.01	0.13	0.02	0.32	0.76	0.00
Student applied to UB in 9th grade	0.45	0.02	0.48	0.04	0.76	0.45	0.06
Student applied to UB in 10th grade	0.33	0.02	0.30	0.03	0.91	0.36	0.07
Student applied to UB in 11th grade	0.09	0.01	0.09	0.02	0.37	0.71	0.00

Table notes: Asterisks indicate that the mean differences are statistically significant at p= .05 or lower, a condition to be addressed in the estimation phase. Multiplying each student in the sample by the inverse of their selection probability for a given stratum and then adjusting these stratification weights to reflect population level characteristics created post-stratification weights.

[1] Standardized mean differences. Calculated by dividing the absolute difference in the group means by the treatment group standard deviation (Stuart, 2010).

RSE= Robust Standard Errors

+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Table 3.6 shows balance test results using the non-response weights that were utilized by MPR researchers to estimate the treatment impact on post-secondary outcomes. In contrast to the prior tables, the sample appears to be well balanced on all variables and this balance is evidently a function of the non-response weights. To reiterate, the Horizons study used a

complex sampling design (first randomly selecting projects and then randomly selecting students) to generate the sample. These weights act to correct for the different probabilities of selecting an individual student caused by this design through weighting each student by the inverse probability of their selection multiplied by the inverse probability of response, which seems to improve balance, and also modifies the variance structure to account for design effects. However, the Horizons non-response weights were constructed to give non-zero weights only those students who responded to the third survey wave for high school graduation outcomes and give non-zero weights only to those students who responded to the fifth survey wave for post-secondary outcomes. Applying these weights would cause hundreds of responses in other survey waves to be excluded from the analysis (see table 3.3). This is potentially unadvisable. The fact that I observe imbalances in the tables where post-stratification or no weights are used, and do not observe imbalances in the table where non-response weights are used suggests that balance is associated with non-response. [17]

[17] Global F-tests for sample balance using the joint distributions of pretest variables under the 3 weighting scenarios did not find evidence of treatment and control imbalances using the $\alpha = 0.05$ threshold. However as Stuart (2010) notes, hypothesis tests that utilize sample size information (of which an F-test is one) should not be used as the sole measure of balance.

Table 3.6. Balance Test Results for the Horizons Study Sample – Using Non-Response Weights

	Control		Treatment		Differences		
Variable	M	RSE	M	RSE	t	p	SMD [1]
Student Educational Expectations	8.24	0.16	8.07	0.18	0.44	0.67	0.03
Low Income Household	0.82	0.02	0.80	0.04	0.87	0.38	0.01
Quarter of Birth	71.93	0.24	71.26	0.42	1.57	0.12	0.04
Female	0.71	0.03	0.71	0.04	0.96	0.33	0.00
White	0.19	0.02	0.03	0.03	0.95	0.33	0.14
Hispanic	0.23	0.03	0.23	0.04	0.73	0.89	0.00
Black %	0.52	0.03	0.49	0.04	0.62	0.53	0.02
Other race	0.06	0.01	0.07	0.01	0.69	0.49	0.03
At least one parent has a B.A. or higher	0.04	0.00	0.04	0.01	0.41	0.68	0.00
Project Director expects to serve this student	1.62	0.03	1.63	0.05	0.20	0.85	0.01
Student previously received Precollege services	0.32	0.03	0.31	0.04	0.10	0.93	0.01
Student is attending the same school as last year	0.60	0.03	0.59	0.04	0.10	0.93	0.01
Student is living at the same address as last school year	0.83	0.02	0.82	0.03	0.17	0.86	0.01
Student applied to UB in 8th grade	0.14	0.01	0.13	0.02	0.75	0.45	0.01
Student applied to UB in 9th grade	0.45	0.03	0.44	0.04	0.14	0.89	0.01
Student applied to UB in 10th grade	0.33	0.03	0.33	0.04	0.35	0.73	0.00
Student applied to UB in 11th grade	0.08	0.01	0.10	0.03	0.89	0.37	0.02

Table notes: MPR researchers created non-response weights by adjusting the post-stratification weights to reflect survey attrition and permit generalization. [1] Standardized mean differences. Calculated by dividing the absolute difference in group means by the treatment group standard deviation (Stuart, 2010)
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Missing Data

Approximately 20% of students assigned to treatment never showed up for any program activities. The existence of "no-shows" could indicate the potential confounding of treatment receipt with educational outcomes. Students who were selected for treatment but were "no shows" were asked in a follow up interview why they did not enroll in UB. The major reasons given (multiple responses allowed) were transportation problems (20% of respondents), they

took a job (18%) or they were never contacted (18%). Also, "no show" rates among students that were considered to be "at risk" (unspecified) were 25% versus 20% for the entire sample (Myers and Schirm, 1999).

As is common in longitudinal studies, not all students who participated in some program activities completed all years of the intervention. Roughly 35% of the treatment group completed all years of UB for which they were eligible (Myers and Schirm, 1999). The grantees surveyed for the retrospective study of UB reported similar persistence figures for their students as were reported by the Horizons study team suggesting that the level of attrition the Horizons study researchers encountered was not unusual (Myers and Moore, 1997).

Response rates for each round of survey declined over time. Table 3.7 shows that approximately 26% of the original sample had quit answering the surveys by the 5th round. Table 3.7 also highlights the incidence of differential attrition between treatment and control samples. Starting with the 2nd follow up survey treatment group responses were five percentage points higher than control group responses (88% versus 83%) and this difference persisted through the study termination date (76% versus 72%). This differential attrition represents a potential threat to internal validity (Shadish, Cook, and Campbell, 2002).

The lower portion of table 3.7 highlights the potential advantages of utilizing multiple survey wave data in estimating program effects. The differential attrition when using cumulative responses is smaller for two of the outcome measures, which reduces the potential threat to internal validity. The overall survey attrition, and therefore the amount of missing data are smaller for all outcome measures, which reduce the number of dropped cases or reliance on external data to fill in for missing data.

MPR researchers looked to external data sources to address missing post-secondary outcome data.[18] Missing outcome measures for 5th follow up survey responders were replaced with data collected from two outside sources, the NSC and federal Student Financial Aid (SFA) records (Seftor, et al., 2009; Cahalan, 2009). The Office of Policy and Program Studies Service did not include the raw NSC data in the release to me that were used by COE and the Horizons research teams in their analyses (PPSS, 2011). [19]

Table 3.7. Response Rates By Survey Wave

Single Survey Wave	Percent Responding (Initial Sample = 100)		
	Total Sample	Treatment	Control
Baseline (1992-1993)	99	99	99
First Follow Up (1994-1995)	97	97	96
Second Follow Up (1996-1997)	86	88	83
Third Follow Up (1998-1999)	81	83	78
Fourth Follow Up (2001-2002)	75	78	72
Fifth Follow Up (2003-2004)	74	76	72
Cumulative Survey Waves			
Cumulative responses for high school graduation (Second through Fourth Follow Up)	93	94	91
Cumulative responses for Post-secondary enrollment (Third through Fifth Follow Up)	94	95	92
Cumulative responses for Post-secondary completion (Third through Fifth Follow Up)	89	90	86

Sources: Seftor, N. S., Mamun, A., and Schirm, A. (2009). PPSS (2011).

[18] Adelman (1999) notes that ambiguity in how college enrollment survey questions were asked in self-administered surveys can introduces high levels of volatility in year-to-year enrollment rates. Administrative databases such as those put together by the American College Testing Service have higher accuracy while being much less volatile.

[19] Horizons researchers noted that under coverage might limit the usefulness of the NSC data and that measurement error might limit the usefulness of the SFA data, but did not appear to quantify the size of the limitation (Seftor, et al., 2009).

In addition to testing for balance between treatment and control groups, I also tested for differences in baseline characteristics between those who reported their outcome measures and those who did not. Statistically significant differences in baselines between those who stay in the study and those who leave might indicate a threat to internal validity, but this threat can be mitigated thorough covariate adjustments in the estimation equations. All of the following reported differences were found to be statistically significant at p= .05.

A higher percent of students with missing high school outcomes were from a low-income household (12 percentage points), likely to be served by UB (20 percentage points), male (17 percentage points), non-white (7 percentage points), or had not previously received precollege services (12 percentage points). Higher percentage points of students with missing post-secondary outcomes were male (26 for enrollment, 23 for completion), non-white (10, 7), applied to UB in the 9th grade (18, 11) or had not previously received precollege services (13, 9). Also student who applied to UB in the 10th grade had a lesser proportion of missing post-secondary outcomes (12, 11).

The Horizons method for solving the problem of missing data was to drop those cases that did not respond to key survey covariates and adjust the non-response weights accordingly. My approach was different. I used multiple imputation with deletion (MID) to gauge how robust my results were to missing covariate data (von Hippel, 2007).

Incidence of missing baseline data was minimal. The only important pretest variable for which more than 39 data observations were missing was student educational expectations with 246 missing observations. This variable had a relatively high amount of missing data because

215 responses were given as Don't Know. I recoded these uninformative responses as missing and then imputed.[20]

Allison (2002) explains that missing data can be one of three types, *missing completely at random* (MCAR), *missing at random* (MAR), or *not missing at random* (NMAR). As is generally the case, I cannot assert that the data are MCAR and so by default they are considered MAR, which would seem to be a reasonable classification of the data because I am only imputing data that were collected pre-randomization.

Attrition represents a form of potential selection bias because it occurs after treatment assignment (Shadish, Cook, and Campbell, 2002). However I can act to address this problem by using a covariate adjustment model. Separately, I identify the size of the potential problem caused by missing outcomes by testing the sensitivity of my results to missing outcome data using the bounding strategy pioneered by Manski (Manski, 1995; Manski, 2003; Morgan and Winship, 2007). Notably, as the proportion of the missing outcome data increases, the width of the bounds also increases.

Imputation

Multiple imputation with deletion (MID) is increasingly recognized as an effective strategy when data are classified as MAR (Von Hippel, 2007). The MID strategy involves imputing the missing data and then deleting the imputed observations that contained previously missing Y data. This second step needs to be undertaken for each outcome measure. Post-deletion, I analyzed the recombined data set using the same OLS estimation models as used under listwise deletion (Von Hippel, 2007).

[20] I considered using a "don't know" indicator variable in the OLS regression equations but rejected this approach as it might introduce bias (Jones, 1996; Allison, 2009)

The rationale for including the dependent variable as part of the chained equations that impute the missing independent variables is intuitive. I took this step to acknowledge the relationship between X and Y. If I left Y out of the imputation process, the imputed X variables have no relationship to Y except by chance. The end result is that the relationship between the imputed X variables and the existing Y (in the absence of using Y to impute the missing X's) are biased towards zero (Allison, 2002;Von Hippel, 2007). I took the next step, deleting those cases with imputed Y's (irrespective of whether these same cases also have imputed X's) because those cases did not inform my understanding of the relationship between the X's and Y and only added estimation error to the regression equations (Von Hippel, 2007). The outcome of the MID process is the imputation of missing X variables for cases where the Y variable is known.

Manski Bounds

To address missing outcome data, Manksi's partial identification strategy ("Manski Bounds") can be employed to detect average treatment effects (ATE) when working with incomplete data (Manski, 1995; Manski, 2003; Morgan and Winship, 2007). Manski's approach is to impute values for those cases with missing outcomes (Manski, 1995; Morgan and Winship, 2007). Under the "no assumptions" conditions, the ATE is bounded on the low end by minus one and on the high end by plus one for dichotomous potential outcomes, such as the ones I examined in this dissertation (Morgan and Winship, 2007).

The universe of possible solutions can be further defined by invoking two weak, yet plausible assumptions, Monotonic Treatment Response (MTR) and Monotonic Treatment Selection (MTS). MTR assumes that the treatment effect cannot be negative, which sets the lower bound ATE at zero. Put another way, under MTR control group member outcomes are not made worse by exposure to treatment. MTS can be used to tighten the bounds because it implies

that treatment recipients are expected to have higher average outcomes than participants assigned to treatment who do not receive it (Morgan and Winship, 2007). In my sensitivity analyses I used bounds produced by applying the MTR and MTS assumptions.

Sample Non-representativeness

The Horizons study design required that each UB site in the experiment have at least twice as many applicants as there were open slots (Myers and Schirm, 1999). This requirement appears to have caused the site directors to modify their standard student selection criteria. A national review of the operational practices at UB sites and experiences of UB feeder schools during the 1990's (i.e., the grantee survey) detailed a potentially different set of norms used by site directors to identify eligible students (Moore, Fasciano, Jacobson, Myers, and Waldman, 1997).

The grantee survey, also conducted by MPR, was given to a nationally representative sample of 244 grantees, 220 of which filled out the survey. Survey respondents included the 67 grantees randomly chosen for inclusion in the Horizons study sample (Moore, et al., 1997).[21]

Local UB site directors were asked as part of the grantee survey to disclose any and all selection criteria routinely used to determine student eligibility for that director's program. Site directors reported that they regularly used a two-step process to construct their list of eligible candidates once federal eligibility had been established (Moore, et al., 1997). As a first step, 95% of project directors relied on positive recommendations from feeder school educators and staff before considering a student eligible to receive local UB services. The Horizons study reports

[21] The universe from which those 244 grantees were selected was a national sampling frame of 440 grantees that had been in operation for more than three years, and were chartered as a two or four year college or university (Moore, et al., 1997).

document no such step in their design (Myers and Schirm 1997; Myers and Schirm, 1999: Myers, et al., 2004; Seftor, et al., 2009).

Once the initial pool of students had been established, 86% of project directors then used a second step, utilizing one or more screening criteria to identify ineligible students.[22] The top seven reasons used by site directors to render a student ineligible were (Moore, et al., 1997):

1) No specific interest in college (46%)
2) History of emotional or behavioral problems (41%)
3) History of drug or alcohol abuse (41%)
4) Gang activity (34%)
5) Record of disciplinary actions (34%)
6) GPA above a specified minimum (32%)
7) GPA below a specified minimum (26%)

Application of one or more of these criteria has the potential to radically reduce the size of the applicant pool. For instance, 62% of site directors immediately excluded any student who reported any history of behavioral problems (Moore et al., 1997). However, again, none of these seven criteria were apparently in force for the recruiting phase of the Horizons study (Myers and Schirm 1997; Myers and Schirm, 1999: Myers, Olsen, Seftor, Young, and Tuttle 2004; Seftor, Mamun, and Schirm 2009). These differences in recruiting practices suggest differences in characteristics between the Horizons students and typical UB students.

In support of the idea that the Horizon study student sample is different that the typical UB student population I offer the following data. In conjunction with the grantee survey report MPR also collected survey data on the experiences of the UB feeder schools. A nationally representative sample of 550 school staff that acted as liaisons to UB were asked what percentage of their applicants had been rejected by UB over the last three years (Moore, et al.,

[22] Questions C5 and C6 from the grantee survey (Moore, et al., 1997).

1997). A total of 222 liaisons (40.4% of the sample) reported that more than half of their applicants were routinely rejected by UB. A weighted average of the responses of all 550 liaisons showed the overall applicant exclusion rate for the three years that preceded the Horizons study to be roughly 43.4% (Moore, et al., 1997). This exclusion rate contrasts with an apparent Horizons study exclusion rate of zero.

A second observable indicator of sample differences comes from the answers to a grantee survey question that queried site directors about their recruiting practices (Moore, et al., 1997). [23] Historically, about one-third of site directors pre-screened the applicant pool by either targeting their recruiting efforts at those students who were most likely to meet that project's eligibility requirements or by recruiting just enough eligible students to fill the number of expected project openings (Moore, et al., 1997). The remaining two-thirds of project directors accepted applications from all interested students, but then applied screening criteria to sort out those students they considered to be ineligible. However, the RCT design used in the Horizons study mandated that all site directors recruit at least two students for each opening to support random assignment (Myers and Schirm, 1999). This difference in recruiting practices employed for the experiment suggests that at least some directors were compelled to over-recruit to fulfill the requirements of the experiment (Cahalan, 2009). This potential issue was apparently not considered a problem by Horizon researchers, nor was its potential effects on the composition of the study sample addressed in the Horizon study reports or the COE report.

The same data that indicate differences in the sample due to over-recruitment and altered selection processes also provide a way to account for the problem. Motivated by the lack of

[23] Question B8 from the target school survey (Moore, et al., 1997).

correlation between the director's assessment of the eligibility of an individual student (as noted on the student eligibility form) and the listed eligibility criteria as published by Moore, et al. (1997), I built a screening tool using student responses from the Horizons study baseline survey (administered pre-intervention) to differentiate between students who might normally be considered eligible and those who might be classified as ineligible once the screening criteria were applied. Put another way, I was able to utilize the Horizons study data to potentially identify those students who would have been most and least likely to participate in UB under typical implementation.

The complete table of survey questions and threshold responses is displayed in table 3.8. In building my screening variable, I classified as ineligible all students whose response to any single question met or exceeded the threshold numbers I set. For example a student who reported being late for school 10 or more times during the last school year, would be classified as ineligible, as would a student who was put in juvenile detention. My threshold levels were informed by the grantee survey.

Table 3.8. Screening Tool Questions

Survey Question	Threshold Response	Response Range
During the last school year, how often did each of the following things happen to you?		
I was late for school	10 or more times	0- 10 or more times
I cut or skipped classes	10 or more times	0- 10 or more times
I missed a day of school	10 or more times	0- 10 or more times
I got in trouble for not following school rules	10 or more times	0- 10 or more times
I was put on in-school suspension	3 or more times	0- 10 or more times
I was suspended or put on probation from school	1 or more times	0- 10 or more times
I was transferred to another school for disciplinary reasons	1 or more times	0- 10 or more times
I was arrested	1 or more times	0- 10 or more times
I spent time in a juvenile home/detention center	1 or more times	0- 10 or more times
How sure are you that you will graduate from high school?	I probably won't graduate	Probably won't- Definitely will
How far in school do you think you will go?	Finish College	Won't finish high school- Earn PhD

Table notes: Questions taken from the Horizons Study Baseline Survey, administered between 1992 and 1994. Threshold levels informed by UB director's responses to behavioral and academic questions posed in Moore et al., (1997).

A total of 976 out of 2820 students who filled out the baseline survey (34.6%) would have been classified as ineligible using such a screening tool. This ad hoc process does not take into account screening criteria that were absent from the baseline survey data but were typically used by site directors to exclude applicants such as a poor recommendation from the feeder school, drug or alcohol abuse, gang involvement or GPAs that were too high or too low (Moore, et al., 1997). Therefore my screening filter could be considered too coarse because I am not screening out a high enough percentage of applicants. My 34.6% ineligible rate is almost 10 percentage points lower than the 43.4 % exclusion rate experienced in the three years prior to the Horizons experiment by the UB feeder schools (Moore, et al., 1997).

Probability Weights

There are two apparent problems with the probability weights: First, if the goal is to identify estimates that relate to the larger UB population under natural implementation procedures, then a different set of post-stratification and non-response weights would have to be calculated just for the eligible population. (The sampling weights are unaffected except if a site has no typically eligible students.) Second, there is an unusually large range of weights, which may unnecessarily inflate the variance. I discuss these in turn.

There were numerous differences between eligible and ineligible students, and ineligible students ostensibly made up between 34.6% and 43.4% of the Horizons sample. This observation created a dilemma. Neither set of weights (i.e., non-response or post-stratification) can accurately describe the distribution of in-school and out-of school behavioral characteristics of this group as those characteristics are purposely omitted from the general UB population. On the other hand, not using weights might limit generalizability. Put another way, problems with the sample made the use of probability weights less effective in establishing external validity because the best-case scenario is that I was generalizing to a sample that was already non-representative of the population of students actually selected for UB.

A second problem is that there is unusually large variation in the weights across sites. In general, such variability can lead to such large increases in the variance of effect estimates that trimming the weights and allowing some bias may be preferable (Potter, 1988; Potter, 1990; Henry and Valliant, 2012). There is a significant literature on the use of weight trimming strategies in the face of extreme sampling weights (Potter, 1988; Potter, 1990; Liu, Ferraro, Wilson, and Brick, 2004; Pedlow, Wang, Yongyi, Scheib, and Shin, 2005; Chowdhury, Khare, and Wolter, 2007; Elliot, 2008). The need for trimming is only unambiguous, however, when the

weights are ignorable, meaning that they have little influence on the point estimates (Asparouhov, and Muthen, 2005). Research areas in which trimming is used include studies conducted or sponsored by the National Institute of Health, National Assessment of Educational Progress, the Bureau of Transportation Statistics, the Centers for Disease Control, to name a few. The trade-off between variance and bias is often measured by the mean-squared error (MSE) and some simulation evidence suggests that trimming weights at the 30th and 75th percentiles yields the smallest MSE (Asparouhov and Muthen, 2005)).

MPR researchers conducted numerous weight sensitivity analyses in the final evaluation report. The purpose of these analyses was to identify the impact of the sampling weights for project 69 on the point estimates and standard errors. As Seftor, et al. (2009) stated:

> This Appendix examines the extent to which Project 69 is unusual, how impacts would change if the weights were different, and the role that Project 69 plays in affecting both the point estimates and the standard errors of the estimates. (p. G.4)

To answer those questions, MPR researchers collapsed the number of strata by reducing the number of stratifying variables, re-stratified the sample based upon distance measures of combinations of student and project characteristics, redistributed the sampling weights to other projects to reduce the weight of project 69, reweighted the projects based upon project size rather than number of federally eligible applicants, assumed no weights, and solved for the effect sizes that project 69 would have needed to have been given to make the treatment statistically significant in the weighted estimation equations. Through these scenarios MPR researchers were able to generate cases where UB was estimated to be effective in increasing post-secondary enrollment and completion. In the end however, the report recommended staying with the

weighting structure that was originally chosen. The reasons given were that project 69 was not unusual; that changing weights would reduce generalizability, and that there was no policy context for interpreting the findings that arose from using different weights (Seftor et al., 2009).

In contrast, following the above research literature, I considered the question of whether the sampling weights in the Horizons study were extreme in ways that might significantly increase the mean squared error of the estimates. Students in five of the 67 projects, which represented three of the 46 strata, were assigned approximately 43 percent of the weighted sample (Seftor et al., (2009). As an example, the 85 students in project 69, who were 3% of the non-weighted sample, made up 26% of the post-stratification weighted sample. Across all students in the study, sampling weights ranged from 1.0 to 79.6, with a mean of 7.2, and a standard deviation of 10.9. Post-stratification weights ranged from 1.9 to 184.8, with a mean of 15.4 and a standard deviation of 24.1 (PPSS, 2011). These large variations in sampling weights present a problem for the analysis of survey data because this variation greatly increases the variability of the point estimates (Asparouhov and Muthen, 2005). For my analysis I trimmed the post-stratification weights at the 75th percentile. I chose the post-stratification sampling weights for the following reasons. First, simulations run by Asparouhov and Muthen support the idea that trimming reduces the MSE. Second, Seftor et al., (2009) made the point that the nonresponsive adjustment portion of the weight is small compared to the portion due to the selection of the project site.[24] Third, the post-stratification weights are, of the available weight choices, the ones best suited to support claims of external validity

[24] Page G.16 of the final report

In addition I tested the sensitivity of the estimated effects without weights, while accounting for site clustering. This method would arguably still allow for establishing internal validity (Elliot, 2008). Coupled with the effect heterogeneity strategy recommended above, this approach represents one approach to a difficult set of problems.

Replication of Horizons and COE Findings

I started by replicating the Horizons findings for the effect of treatment on 12th grade GPA, high school graduation rates, and post-secondary enrollment and completion rates, to the extent permitted by the released data. I acquired the data and the documented equations required to replicate 12th grade GPA, and high school graduation rates, but as previously mentioned I lacked the data to perfectly duplicate the post-secondary outcomes featured in the Horizons evaluation reports.

Since I was unable to precisely replicate the featured post-secondary completion analyses (because data for that were not provided) I instead replicated an ancillary set of analyses that the MPR researchers conducted using only student survey responses (i.e., devoid of NSC data and SFA financial aid applicants who did not receive a Pell grant).[25] These ancillary analyses were distinguished from the featured analyses in that the former suggested a positive significant treatment effect on post-secondary completion, while the SFA and NSC-based estimates suggested no such effect.

Next I replicated COE researchers findings using their assumptions, models, and outcome measures, which differed from those used in the Horizons study. When estimating the effects of UB on post-secondary outcomes COE standardized the outcome measures by keying these

[25] These findings were published in the 5th follow up report in Appendix C, as Table C.7 Case #1 and Table C.14 Case #1 (Seftor, Mamun, and Schirm 2009).

measures to the expected high school graduation year for a given student (e.g., post-secondary enrollment within 18 months of expected high school graduation year, and B.A. attainment within 6 years of expected high school graduation year) and then calculated treatment effects on those student groups. COE used SFA data to fill in for missing post-secondary outcomes. COE did not analyze the effects of UB on high school outcomes.

Constructing New Effect Estimates

I constructed estimates using methods that are at least equally reasonable if not preferable for the four outcomes as measured by MPR researchers. The methods I used to develop my final data set and produce my estimation equations differ from those used by MPR researchers in six ways. The first four differences addressed internal validity issues; the last two primarily addressed external validity issues.

To reiterate, I used all available student responses, across all relevant survey waves. First, for high school outcomes, I used the data captured by survey waves 2-4, while MPR researchers used only survey wave 3 data. Second, for post-secondary outcomes, I used the data captured by survey waves 3-5 while MPR researchers used survey wave 5 data. Using more data waves allowed me to add several hundred more valid responses to the data set. Third, my data set contains incomplete supplemental post-secondary enrollment and completion data from the National Student Clearinghouse (NSC) and Student Financial Aid (SFA) (PPSS, 2011)

A fourth difference is that I identified two variables, student educational expectations and low income that were out of balance in some specifications. These two variables are associated with educational outcomes and I added these as covariates in my regression equations to pseudo-balance them (Alexander, Entwisle and Horsey, 1997; Rumberger and Lim, 2008). In contrast,

MPR researchers used a complex sampling design and associated post-stratification and non-response weights to account for these problems and did not include covariates when estimating high school outcomes.

A fifth difference is that I estimate effects with various alternative weights including trimmed weights, as described in the previous section. Incorporating either post-stratification or non-response weights into my estimation equations rendered my effect estimates for all outcomes as insignificant. So that my estimates still accounted for site-level differences in errors in the absence of survey weights, I clustered my errors by site.

The sixth difference is that I accounted for differences in site director selection criteria between experimental and non-experimental years and demonstrated how those differences produced a sample that was different from the national UB program. I used the data to explore the possibility of effect heterogeneity by eligibility.

I subjected my findings to a series of sensitivity tests to establish the robustness of my ITT results. The sensitivity tests cover a number of different scenarios. I examined how robust my estimates are to missing covariate observations by imputing the missing data and also to missing outcome by bounding the problem using an approach pioneered by Manski (1995). Finally, I used instrumental variables to quantify how much my effect estimates change when I instead calculate the impact of treatment receipt (TOT) on educational outcomes.

Estimation Techniques

I executed a series of regression equations to estimate the effects of treatment assignment (ITT) and estimate the impact of treatment receipt (TOT) on each of the four educational

outcomes. I give a complete explanation of the TOT rationale and approach in the section following this one.

The treatment effect for the unadjusted model was estimated using the following regression equation:

$$(1) \qquad Y_i = \beta_0 + \beta_1 T_i + \varepsilon_i$$

Where Y is the educational outcome of interest, T indicates the treatment assignment (implying that these are ITT estimates), and the "i" subscript indicates that treatment occurs at the individual level.

To replicate Horizons study findings for the effects of UB on high school graduation rates and 12th grade GPA's I used equation (1) with appropriate non-response weights. These equations duplicated the ones referenced by the Horizons study (Myers et al., 2004). COE did not estimate the effects of UB on high school outcomes.

Regression-based covariate adjustments can increase power when estimating treatment effects.[26] Since I encountered treatment and control group imbalances on key background variables, I used covariate adjustment models to address imbalances. The covariate-adjusted ITT effect was estimated using the following regression equation:

$$(2) \qquad Y_i = \beta_0 + \beta_1 T_i + X_i \Gamma + \varepsilon_i$$

Where Y is the educational outcome of interest, T indicates the treatment condition, and the "i" subscript indicates that treatment occurs at the individual level. I used X to represent a vector of other pretest covariates (including race, gender and grade at time of application). All

[26] Some researchers argue against the practice of using regression adjustment models when analyzing experimental data because the post-adjustment standard errors may be overly large or small (for example, see Freedman (2008)). However this objection does not appear to apply to experiments with sample sizes of over 1000 (as noted by Freedman (2008) on page 191).

other variables in the sample appear to be well balanced between the treatment and control groups.

To replicate Horizons study ITT findings for the effects of UB on post-secondary enrollment and completion, I used equation (2) with appropriate non-response weights, plus interactions between the covariate vector and an indicator for project 69. These equations duplicate the ones referenced by MPR researchers (Seftor et al., 2009). COE estimated the ITT effects of UB on post-secondary enrollment and B.A. completion. To isolate the contributions of project 69 students to the effect estimates, COE conducted their analyses with and without these students.

To replicate COE's findings, I used equation (2) in conjunction with Stata's survey command set, which recognized the stratified nature of the data, allowed for the use of probability weights, and also clustered the data into defined groups to generate the necessary robust standard errors. COE used post-stratification weights. For COE's analyses excluding project 69, the number of strata was reduced from 28 to 27 and the number of defined groups was reduced from 67 to 66.

Effect heterogeneity is present when the causal effect of treatment is different across individuals (Elwert and Winship, 2010). Evidence of effect heterogeneity is strong when three conditions are met (Bloom and Michalopoulos, 2010).[27] The first condition is that the ATE is statistically significant for the full study sample. The second condition is that the intervention is shown to be statistically significant for the identified subgroup(s). The third condition is that

[27] Not all researchers agree with requiring an average treatment effect to be non-zero (for example see Harris and Goldrick-Rab, 2012)

differences in effect estimates between the identified subgroup and the other sample group members are statistically significant.

One test for meeting the third condition is to use an OLS regression equation (with clustered standard errors) to measure the interaction effect between the identified subgroup and the treatment variable for a given educational outcome as shown in equation (3).

$$(3) \qquad Y_i = \beta_0 + \beta_1 T_i + \beta_2 S_i + \beta_3 T_i \cdot S_i + X_i \Gamma + X_i \Gamma \cdot S_i + \varepsilon_i$$

Where Y is the educational outcome of interest, T indicates the treatment condition, and the "i" subscript indicates that treatment occurs at the individual level. I used X to represent a vector of other pretest covariates (including race, gender and grade at time of application). S is the student subgroup and $T \cdot S$ and $X_i \Gamma \cdot S_i$ represent the interaction terms. Evidence of a subgroup difference is presented as a statistically significant coefficient for the $T \cdot S$ interaction term.

One test for meeting the second condition is to fully stratify (3) on the subgroups of interest. Stratification supports measuring the effect of treatment on the identified subgroup.

The pretest variables I have identified which are shown in the literature or were hypothesized by MPR researchers to affect student outcomes are: taking algebra or above in the 9th grade, a 9th grade GPA over 2.5, and student educational expectations of a B.A. or above (Myers and Schirm, 1999; Adelman, 2006: Rumberger and Lim, 2008). Consistent with the practices of other researchers, I named these variables before analyzing the data to avoid any possibility of "data-mining" (Imai and Ratkovic, 2010; Bloom and Michalopoulos 2010). In addition I tested for effect heterogeneity by eligibility (Moore et al., 1997).

Balance tests conducted on the four subgroups (i.e., 9th grade GPA was above or below 2.5, student expects to receive a B.A. or above, student did/did not take Algebra in the 9th grade,

and local program eligibility of the student) showed that only student educational expectations of a B.A. was correlated with treatment assignment, and the correlation was negative. Therefore the treatment effect estimated for this subgroup likely understated the true effect of treatment.

I also conducted an exploratory analysis to investigate whether the impact estimates for high school graduation, post-secondary enrollment, and post-secondary completion were statistically significant across UB sites. Testing for impact variation across sites is an emerging field in education research (Bloom, 2012). To test for the possible presence of impact variation I interacted treatment with site, and included the vector of controlling covariates, as well as the treatment variable.

Instrumental Variables

The analyses conducted by MPR researchers and COE produced estimated effects of treatment assignment (ITT) and estimated impacts of treatment receipt (TOT). ITT represents estimated effects of treatment on the assigned population. TOT represents the estimated impacts on compliers. I defined treatment receipt as students having attended at least one session of UB or its companion program Upward Bound Math and Science session (N=43).

To address potential selection bias, I use an instrumental variable identification (IV) strategy, where treatment assignment was my instrument to uncover the effects of treatment on those who actually received treatment (Bloom, 2005).[28] For the Horizons experiment approximately 80% of the students assigned to treatment group received treatment. In an experiment the treatment assignment variable can serve as an ideal instrument because in expectation it is independent of all other variables at the point of randomization (Bloom, 2005).

[28] I used Stata command ivregress.

By definition, a valid IV must meet the following two criteria. The instrument "Z" is correlated with "T", treatment receipt (instrument relevance), and "Z" is independent of the error term (instrument exogeneity), an assumption that generally cannot be tested (Morgan and Winship, 2007; Gelman and Hill 2007). By design Horizons students were permitted to attend UB only by assignment, which establishes the relevance of the instrument. Also by design, students were randomly assigned to treatment or control, which makes instrument exogeneity a reasonable assumption, and generally satisfies the exclusion restriction. There are some other substantive assumptions that are not usually mentioned.[29]

I used two-stage least squares method to estimate the TOT. In stage one I regressed the dependent variable on treatment assignment to create an unbiased estimator of treatment receipt. In stage two I used this newly created variable in an OLS equation to estimate treatment impacts.

For UB the TOT being estimated is the complier average causal effect (CACE). Compliers represent the portion of the experimental population who are induced to take or not take treatment based solely upon treatment assignment. By definition, always-takers and never-takers are not susceptible to an inducement to receive or refrain from seeking treatment, and crossovers are similarly immune. Therefore the treatment effect is produced solely by those assigned to treatment that seek treatment and those assigned to control that do not receive treatment.

---

[29] In addition to the exclusion restriction, IV analysis requires three other unverifiable substantive assumptions. First, treatment assignment does not affect the potential educational outcomes of the never-takers. Second, treatment assignment does not affect the potential educational outcomes of the always-takers. It is because there are no observable counterfactual conditions for these two groups that these premises are not apparently testable. The final assumption is known as the Stable Unit Treatment Value Assumption (SUTVA). Under SUTVA, assignment of one individual to treatment does not diminish the quantity available to another assigned individual. Also, under SUTVA the quality of the treatment is independent of the treatment provider (Morgan and Winship, 2007).

Power Analysis

I used the free version of G*Power to approximate if my available sample size was sufficient to detect an effect size of 0.2 standard deviations in magnitude given alpha=0.05 and 1-Beta =0.95 for two randomly assigned groups (Faul, Erdfelder, Lang, and Buchner, 2007). The requisite sample size is 1302, or 651 in each treatment condition. I have 2844 observations in my sample, which is large enough to detect an effect size of 0.1 standard deviations given alpha=0.05 and 1-Beta =0.95 for two randomly assigned groups and so I can be confident that I will not overlook a treatment effect for my stated magnitude of 0.2 standard deviations.

MPR researchers also measured the degree to which students who were associated with the same UB project site (whether they were chosen for the experiment or not) resembled each other on their observables. This measure, known as the Interclass Correlation (ICC) was used by MPR researchers to quantify the degree to which variation in student outcomes can be attributed to characteristics shared by students linked to the same UB project site. MPR researchers found that less than 5% of the variance in student outcomes was attributable to a given site (Myers and Schirm, 1999).[30] These modest ICC levels suggest that student background characteristics do not vary much at the site level. However, given that some variance was detected I will still need to adjust for the effects of clustering on the standard errors produced in the estimation phase.

My ICC calculation showed minimal effects of clustering for graduation and enrollment outcomes (ICC< 0.05) and so I think the G*Power estimate is sufficiently close. As an aside, even in cases where my sample size was reduced because of missing observations, I added

[30] I calculated the intra-class correlations for all study outcomes. I can verify that for all outcomes except for high school GPA, the percentage of variation attributable to group membership ranged from 4 to 6 %. For high school GPA's the number was 12%.

controlling covariates that acted to reduce the sample size required to detect the stated effect (Cohen, 1992).

Chapter 4 - Findings

Replication of Published Results

I compared the baseline characteristics for the released data with those published in the final evaluation report to demonstrate that the data I received was the same as the data used in the Horizons study (table 4.1).

Table 4.1 A Comparison of Baseline Characteristics for the Full Evaluation Sample– Using Post-Stratification Weights

	Control Means		Treatment Means	
Variable	Baseline Survey	Horizons Final Report	Baseline Survey	Horizons Final Report
Female	71	71	68	68
White	20	20	21	20
Hispanic	23	22	22	21
Black	51	50	50	50
Other race	6	7	7	7
Low Income Household	83	83	86	83
Student applied to UB in 8th grade	13	13	13	13
Student applied to UB in 9th grade	45	45	48	48
Student applied to UB in 10th grade	33	33	30	30
Student applied to UB in 11th grade	9	9	9	10

Table Notes: Sources: Horizons Baseline Survey file (PPSS, 2011), Seftor, N. S., Mamun, A., and Schirm, A. (2009). pp. A.20. "Baseline Survey" columns refer to my calculations.

MPR researchers produced estimates for the effect of UB on high school graduation, 12th grade GPA, post-secondary enrollment and completion. I replicated their published results for the estimated effects of treatment assignment (ITT) and the estimated impacts of treatment receipt (TOT) and established that I could reliably achieve those results from the released data

using MPR's methods (table 4.2).[31] The results are qualitatively similar, although because the Horizons study does not report standard errors, I can only compare the point estimates and threshold significance levels. I found levels of statistical significance to be the same in all cases, although the point estimates differed.

Consistent with the Horizons study, I found no evidence of a treatment effect on high school graduation (-2.0 percentage points), 12th grade GPA (-.067 points on a 4 point scale), or post-secondary enrollment (0.2 percentage points). My estimates also showed no statistically significantly different effect (i.e., Project 69 interacted with treatment) of having been a student assigned to the Project 69 UB site and these findings are in accord with MPR's published statements. In addition I found statistically significant evidence of a treatment effect on post-secondary completion for the survey sample excluding NSC or SFA data (11.3 percentage points), and this finding is in line with MPR's estimate of 13.0 percentage points. To the degree the estimated effects are different, they are larger in the Horizons study.

[31] A comparison of Horizon study and release files baseline characteristics can be found in the Appendix, table A2.

Table 4.2. Replicating Horizons Findings for the Effect of Upward Bound Assignment (ITT) on Educational Outcomes

Response Variable	High School Graduation		12th Grade GPA		PSE Enrollment		PSE Completion	
	Replic.	Horizons	Replic.	Horizons	Replic.	Horizons	Replic.	Horizons
Effect Estimate	-0.020	-0.010	-0.067	0.000	0.002	0.016	0.113 *	0.130 *
	(0.027)	n.p.	(0.056)	n.p.	(0.026)	n.p.	(0.044)	n.p.
Project 69 Site					-0.071	n.p.	0.292	n.p.
					(0.200)	n.p.	(0.249)	n.p.
Constant	0.896 ***	0.900	2.300 ***	2.300	0.824 ***	0.812	0.447 ***	0.418
	(0.015)	n.p.	(0.037)	n.p.	(0.046)	n.p.	(0.065)	n.p.
N =	2291	2291	2673	2673	2102	2102	1724	1724

Table notes: Horizons high school outcomes are based on 3rd survey wave responses. Horizons PSE outcomes are based on 5th survey wave responses and do not include SFA or NSC data. PSE = Post-secondary education. "Replic."= Replication. MPR researchers generated three separate series of non-response weights; one for high school outcomes, a second for 12th grade GPA and a third for the two PSE outcomes. I used these same sets of weights in attempting to replicate their findings. For PSE outcomes, I replicated the models used to estimate the treatment effect on students who responded to the 5th follow up survey (i.e., no NSC or SFA data). My replications should be compared to those published in the 5th follow up report-Appendix C Tables C.1 case #1 and C.7 case #1. These findings exclude any SFA or NSC data.
Standard errors are in parentheses.
n.p. = not published
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

I was able to approximate or replicate the TOT point estimates published by MPR researchers for all four outcomes. Since the estimate of ITT effects for the high school outcomes and post-secondary enrollment were not statistically significantly different from zero, the TOT impacts are not statistically significantly different from zero.[32] For the post-secondary completion outcome, which was significant for ITT, I calculated an impact estimate of 15.8

[32] This is because Bloom's correction cannot transform an insignificant ITT estimate to a significant TOT one (Bloom, 1984; Bloom, 2005).

percentage points, which matches the Horizons impact estimate of 15.8 percentage points from Appendix C (Seftor, et al., 2009). [33]

COE produced estimates for post-secondary enrollment and B.A. completion. I replicated COE's published results for the estimated effects of treatment assignment (ITT) and the estimated impacts of treatment receipt (TOT) to establish that those results are reliably achievable from the released data using COE's methods (table 4.3).

Like COE, I found UB to have statistically significant effects on post-secondary enrollment for ITT and TOT. Specifically I calculated an effect estimate of 7.5 percentage points versus 6.9 percentage points for COE and an impact estimate of 10.9 percentage points which matches COE's impact estimate. Also, like COE I did not find evidence of a statistically significant effect of treatment on BA completion for ITT or TOT (1.49 and 2.15 percentage points respectively). These estimates are larger by about 0.50 percentage points than COE's published results of 0.90 and 1.70 percentage points.

[33] Table C.14 Case #1 (Seftor, Mamun, and Schirm 2009).

Table 4.3. Recreating The Council on Economic Opportunity's Findings for the Effect of Upward Bound Assignment (ITT) and Enrollment (TOT) on Educational Outcomes

	ITT				TOT			
Response Variable	PSE Enrollment		BA Completion		PSE Enrollment		BA Completion	
	Replic.	COE	Replic.	COE	Replic.	COE	Replic.	COE
Effect Estimate	0.075 ***	0.069 ***	0.015	0.009	0.109 **	0.109 ***	0.022	0.017
	(0.028)	(0.022)	(0.026)	n.p.	(0.040)	n.p.	(0.037)	n.p.
Constant	0.704 ***	0.660	0.243 ***	0.160	0.699 ***	0.625	0.242 ***	0.174
	(0.086)	(0.347)	(0.065)	n.p.	(0.094)	n.p.	(0.068)	n.p.
N =	2813	2813	2813	2813	2813	2813	2813	2813

Table notes: COE data comes from table 5 (p.36) and table 10 (p.43) (Cahalan, 2009). COE response variables use student responses from survey waves 3-5. COE used standardized outcome measures based on baseline survey question B1 (expected high school graduation year) with a correction for 1991-92 responders. COE used the Stata survey command set to generate their results. Number of strata (wprstco)= 28; Number of PSU (wprojid) = 67. Weights used were the post-stratification baseline weights (v5bwgtp1). Weights were unadjusted for non-response. SFA data are included in COE data. PSE = Post-secondary education. "Replic."= Replication.
Standard errors in parentheses
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$
n.p. = not published

Dropping project 69 from the analysis sample increased the effect of UB on post-secondary enrollment for ITT (2.2 percentage points) and TOT (3.1 to 3.3 percentage points) and recast the treatment effects on BA completion for ITT and TOT as statistically significant (table 4.4). These results approximated those found by COE. My post-secondary enrollment calculations showed an effect estimate of 9.7 percentage points and an impact estimate of 14.0 percentage points versus COE's effect estimate of 9.1 percentage points and impact estimate of 14.2 percentage points. My BA completion calculations showed an effect estimate of 4.8 percentage points and an impact estimate of 6.9 percentage points versus COE's effect estimate of 3.7 percentage points and impact estimate of 7.0 percentage points.

The magnitude of the changes in effect estimates brought about by deleting the project 69 students was large. I calculated that post-secondary enrollment and BA completion rates increased by approximately 30% as a result of the change in sample composition. The control group mean outcomes (i.e. the constant terms) were consistently lower for the population excluding project 69, and the relative decline (on the order of 40-50%) is greatest for the BA completion columns. COE argued one reason to drop project 69 is those students were less academically inclined; a line of reasoning apparently not supported by the data.

Table 4.4. Replicating COE's Findings for the Effect of Upward Bound Assignment (ITT) and Enrollment (TOT) After Dropping Project 69

	ITT				TOT			
Response Variable	PSE Enrollment		BA Completion		PSE Enrollment		BA Completion	
	Replic.	COE	Replic.	COE	Replic.	COE	Replic.	COE
Effect Estimate	0.097 **	0.091 ***	0.048***	0.037***	0.140 ***	0.142 ***	0.069***	0.070*
	(0.028)	(0.024)	(0.010)	(0.011)	(0.039)	n.p.	(0.014)	n.p.
Constant	0.622 ***	0.643	0.153 ***	0.133***	0.596 **	0.604	0.141***	0.141
	(0.073)	(0.420)	(0.045)	(0.011)	(0.076)	n.p.	(0.044)	n.p.
N =	2728	2728	2728	2728	2728	2728	2728	2728

Table notes: PSE = Post-secondary enrollment. COE data comes from table 5 (p.36) and table 10 (p.43) (Cahalan, 2009). COE response variables use student responses from survey waves 3-5. The COE B.A. completion outcome includes NSC data. COE response variables use student responses from survey waves 3-5. SFA data is included in the COE PSE enrollment estimate. COE used standardized outcome measures based on baseline survey question B1 (expected high school graduation year) with a correction for 1991-92 responders. COE used the Stata survey command set to generate their results. Number of strata (wprstco)= 27; Number of PSU (wprojid) = 66. Weights used were the post-stratification baseline weights (v5bwgtp1). Weights were unadjusted for non-response. SFA data are included. PSE = Post-secondary education. "Replic."= Replication.
Standard errors in parentheses
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$
n.p. = not published

New Estimates

I produced new effect estimates under the following scenarios: an unadjusted model without weights, a covariate adjustment model without weights, and a covariate adjusted TOT

model without weights. Unweighted models can be used to produce estimates that support internal validity. In addition the estimates produced by these models can be used as a point of comparison when investigating the effects of weight trimming strategies. I tested the sensitivity of my primary model, the covariate adjustment model without weights by imputing missing X values, bounding the missing Y values, utilizing the Horizons study post-stratification and non-response weights and testing for effect heterogeneity. In addition I tested my primary model and the Horizons models for sensitivity to clustering, and trimmed probability weights. I discuss each of these scenarios in turn.

I started by generating unadjusted non-weighted effect estimates for four educational outcomes: high school graduation, 12th grade GPA, post-secondary enrollment and completion (table 4.5).[34] I found evidence that UB assignment had a positive causal effect on high school graduation rates (a 3.4 percentage point increase, results are significant). I did not find evidence of any treatment effect for the other three educational outcomes.

Table 4.5. ITT Impact Estimates for the Unadjusted, Unweighted Model

	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
Unadjusted Model	0.034*	0.044	0.016	0.031
	(0.013)	(0.033)	(0.015)	(0.023)
N =	2645	2674	2660	2517

Clustered standard errors are in parentheses. The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes. Significance levels remained virtually unchanged under logistic regression

$+ p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

[34] I attempted to test for the effect of treatment on educational expectations, but this was difficult because by the time the first post-treatment assignment survey was administered almost 70% of students had already left high school and of the remaining 30% who were still in high school, two-thirds% of these students should have had already graduated.

In order to account for the covariate imbalances I noted from my balance tests, I generated covariate adjusted effect estimates for the four educational outcomes (table 4.6). I considered the covariate adjustment model to be my primary model. I found evidence that UB assignment had a positive causal effect on high school graduation rates (a 4.6 percentage point increase, results are highly significant), post-secondary enrollment rates (a 2.9 percentage point increase, results are modestly significant), and completion rates (a 4.7 percentage point increase, results are significant). I did not find that treatment assignment had a statistically significant effect on 12th grade GPA's.

To put these new estimates in perspective, consider that in the absence of UB approximately 86 out of 100 high school students graduated high school. For students enrolled in UB, that number rose to 91, a relative increase of almost 6%. Comparisons of post-secondary completion rates showed larger relative effects. Absent UB, approximately 43 out of 100 high school students completed their post-secondary education. For students enrolled in UB, that number rose to approximately 48 students, a relative increase of 12%.

Table 4.6. ITT Impact Estimates for the Covariate Adjustment Model

Outcome Measure	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
UB Assignment	0.046***	0.045	0.029+	0.047*
	(0.012)	(0.034)	(0.016)	(0.023)
Student Ed. Exp.	0.026***	0.092***	0.028***	0.022***
	(0.005)	(0.008)	(0.004)	(0.006)
Low Income	-0.075***	-0.043	-0.053**	-0.082**
	(0.016)	(0.044)	(0.016)	(0.027)
Quarter of birth	0.008***	0.034***	0.008***	0.004
	(0.002)	(0.007)	(0.002)	(0.004)
Female	0.03*	0.229***	0.062***	0.071**
	(0.016)	(0.041)	(0.016)	(0.023)
White	-0.054+	-0.138	-0.092**	-0.040
	(0.030)	(0.089)	(0.034)	(0.043)
Hispanic	-0.045	-0.340***	-0.071*	-0.042
	(0.034)	(0.079)	(0.035)	(0.047)
Black	-0.052+	-0.590***	-0.015	-0.016
	(0.027)	(0.085)	(0.028)	(0.042)
Parent(s) have BA	0.093***	0.241**	0.051+	0.086+
	(0.021)	(0.072)	(0.029)	(0.050)
Project Director expects to serve	-0.019+	-0.062	-0.019	-0.020
	(0.011)	(0.039)	(0.011)	(0.020)
Previously received precollege services	0.017	0.043	0.011	0.060**
	(0.013)	(0.038)	(0.017)	(0.021)
School Mobility	0.041+	0.014	-0.0023	0.030
	(0.023)	(0.047)	(0.020)	(0.028)
Address Mobility	0.052*	0.110+	0.031	0.022
	(0.026)	(0.057)	(0.020)	(0.026)
Applied UB-Grade 8	-0.275***	-0.527***	-0.197***	-0.095+
	(0.040)	(0.105)	(0.038)	(0.056)
Applied UB-Grade 9	-0.137***	-0.380***	-0.153***	-0.047
	(0.032)	(0.091)	(0.030)	(0.050)
Applied UB-Grade 10	-0.042+	-0.109	-0.064**	0.027
	(0.022)	(0.067)	(0.020)	(0.043)
Constant	0.228	-0.278	0.179	-0.006
	(0.147)	(0.513)	(0.159)	(0.271)
N=	2400	2428	2414	2285

Table notes: Clustered standard errors in parentheses. The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes. Significance levels remained unchanged under logistic regression

+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

I also investigated the effects of treatment on post-secondary outcomes by types of degrees sought using the covariate adjustment model (table 4.7). The data appears to show that UB increased enrollment at two and four-year institutions. Specifically, I observed an 8.2 percentage point increase in students seeking a two-year or greater degree above the control mean of 78.7%, and I observed an 6.5 percentage point increase in students seeking a four-year or greater degree above the control mean of 44.8%

In addition, UB appears to increase graduation rates among students seeking any degree, or students seeking a B.A., but not those seeking a two-year degree. Specifically, I observed an 4.7 percentage point increase in students seeking any degree above the control mean of 40.8%, and I observed an 3.6 percentage point increase in students seeking a four-year or greater degree above the control mean of 20.8%

Table 4.7. ITT Impact Estimates on Post-secondary Outcomes by Types of Degree Sought for the Covariate Adjustment Model

Effect of UB assignment	Certificates and Above	Two Year Degree and Above	Four Year Degree and Above
PSE Enrollment	0.029^{+}	0.082***	0.065***
	(0.016)	(0.024)	(0.018)
PSE Completion	0.047*	0.016	0.036*
	(0.023)	(0.022)	(0.017)

Table notes: PSE = Post-secondary education. Clustered standard errors are in parentheses. The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes.
$+ p < .10$, $* p < .05$, $** p < .01$, $*** p < .001$

In order to examine treatment impacts I generated TOT impact estimates for all four educational outcomes (table 4.8). I found for students who received treatment, high school graduation rates increased by 5.4 percentage points, post-secondary enrollment rates improved

by 3.4 percentage points, and post-secondary completion rates increased by 5.5 percentage points. All these estimates are statistically significant. These impact estimates are a relative 17% greater than the effect estimates published in table 4.4, and reflect the correction for no-shows (net no-shows were 17%).[35]

Table 4.8. TOT Impact Estimates for the Covariate Adjustment Model

Outcome Measure	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
UB Enrollment	0.054 ***	0.055	0.034 +	0.055*
	(0.014)	(0.041)	(0.019)	(0.027)
Constant	0.232	-0.269	0.184	0.003
	(0.145)	(0.508)	(0.156)	(0.269)
N =	2400	2428	2414	2285

Table notes: PSE = Post-secondary education. The TOT impact estimates were produced using a simultaneous model approximated by 2SLS using Bloom's correction for no-shows (Bloom, 1984;Imbens and Angrist, 1994; Morgan and Winship, 2007). The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes. Stata command is ivregress
Clustered standard errors in parentheses
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

An overall comparative look at the different estimated effects is provided in Table 4.9. I included the sample mean outcomes to give a sense of the magnitude of the effect estimates. In contrast to the Horizons study I estimated a positive statistically significant effect of treatment on high school graduation. In addition I found that UB was modestly effective at increasing post-secondary enrollment rates. COE's post-secondary enrollment effect estimate is 2.4 times greater than mine and is highly significant. Finally, I found that UB is an effective means of

[35] The impact estimate for 12th grade GPA's remained statistically insignificant, as the IV method cannot turn statistically insignificant results into significant ones.

increasing post-secondary completion rates. This finding is consistent with prior research findings by MPR researchers, although their effects are 2.8 times greater than mine.[36]

Table 4.9. A Comparison of New ITT Estimates with Horizons and COE Published Results

	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
My Best ITT Estimate	0.046***	0.045	0.029+	0.047*
	(0.012)	(0.034)	(0.016)	(0.023)
Horizons	-0.010	0.000	0.016	0.130*
	n.p.	n.p.	n.p.	n.p.
COE	N.C.	N.C.	0.069***	0.090
	N.C.	N.C.	(0.130)	n.p.
Sample Mean	0.863	2.418	0.848	0.425

Table Notes: PSE = post-secondary education. The COE measure for PSE completion is B.A. completion and includes NSC data. COE response variables use student responses from survey waves 3-5. SFA data is included in the COE PSE enrollment estimate. The Horizons PSE outcome findings appear in the 5th follow up report-Appendix C Tables C.1 case #1 and C.7 case #1. Horizons high school outcomes are based on 3rd survey wave responses. Horizons PSE outcomes are based on 5th survey wave responses and do not include SFA or NSC data.
N.C. = Not Calculated. COE did not study high school outcomes.
n.p. = not published
Standard errors are in parentheses. My data are clustered to produce the appropriate standard errors. MPR researchers used non-response weights. COE used post-stratification weights
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Sensitivity Analysis

I tested my estimates to measure sensitivity to changes in assumptions. My ITT effects estimates and significance levels were robust to clustering and to missing covariate data (table 4.10). However, these same estimates were highly sensitive to the use of post-stratification or non-response weights in the OLS regression equations. Using either set of weights rendered my findings insignificant. Point estimates and robust standard errors were highly similar for all outcomes except for PSE completion, suggesting that not much new information was gained by

[36] This finding was published in the 5th follow up report-Appendix C. These findings exclude any SFA or NSC data. (Seftor, Mamun, and Schirm 2009).

the use of non-response weights. The post-stratification weighted point estimates for high school graduation and post-secondary completion were similar to the ones produced by the non-weighted (with and without clustering) and imputed covariate adjustment models. The insignificance of these estimates was largely due to the increased standard errors. I explore sensitivity to probability weighting strategies in more detail later.

The unadjusted model only showed evidence of a treatment effect on high school graduation. If I restrict the sample to observations with no missing data on expectations, and then run the unadjusted model I find that differences in point estimates are due in approximately equal parts to changes in the sample and the effect of covariate adjustments.

Manski Bounds

My estimates might also potentially be sensitive to missing Y data. When I imputed the missing outcome data to form the lower bounds and contrasted those findings to the unadjusted model, the effect estimate for high school graduation becomes insignificant. Applying those same magnitude changes to the covariate adjustment model would render the high school and post-secondary completion effect estimates as insignificant.

Table 4.10. Sensitivity Analysis

	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
Covariate Adjustment Model	0.046***	0.045	0.029+	0.047*
	(0.012)	(0.034)	(0.016)	(0.023)
- No Weights	0.046***	0.045	0.029*	0.047*
	(0.013)	(0.029)	(0.014)	(0.021)
- Imputed X Values	0.039**	0.049	0.023+	0.040*
	(0.013)	(0.027)	(0.014)	(0.020)
- Post-stratification weights	0.030	-0.030	0.013	0.046
	(0.024)	(0.052)	(0.025)	(0.038)
- Non-response weights	-0.000	-0.033	0.010	0.065
	(0.024)	(0.052)	(0.027)	(0.045)
Unadjusted Model	0.034*	0.044	0.016	0.031
	(0.013)	(0.033)	(0.015)	(0.023)
- Restricted sample	0.041**	0.037	0.022	0.036
	(0.013)	(0.035)	(0.016)	(0.024)
- Manski [a] Lower Bounds	0.016	N.A.	0.008	0.013
	N.A.	N.A.	N.A.	N.A.

Table notes: PSE= post-secondary education. Restricted sample is for observations that contained student's educational expectations. Clustered standard errors are in parentheses. The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$
[a] Under MTR and MTS assumptions, no weights

Effect Heterogeneity

I conducted tests of effect heterogeneity for each student subgroup and sample remainder as well as the differences between the two for all studied educational outcomes (table 4.11). [37] I found some evidence of effect heterogeneity by eligibility status. I estimated that 12th grade GPA's might increase by 0.12 points on a four-point scale, and post-secondary completion might increase by 8.4 percentage points when contrasted with typically eligible students. The former finding has virtually no practical effect. GPA's would rise from 2.3 to 2.4 on a four-point scale. The latter finding is potentially important, however. The magnitude of the effect for the ineligible student subgroup is a relative 80% larger than I found for the sample as a whole. To put it in perspective, instead of yielding a post-secondary completion percentage of 48 students per 100 UB assignees it would be 56.

Instances where all effect estimates (average treatment effect, subgroup effects and mean differences) are non-zero and statistically significant constitute strong evidence of effect heterogeneity (Bloom and Michalopoulos, 2010). Using that rubric, I found some, albeit not strong, evidence of such an effect.

In addition, I did find statistically significant evidence of cross-site impact variation for each of the outcomes. The impact estimates for high school graduation ranged from 45.5 percentage points above the reference site estimate of 6.3 percentage points to 43.4 percentage points below (i.e., some sites had negative point estimates) with an F-test of 10.69. Post-secondary enrollment impact estimates varied from 12.1 percentage points above the reference site estimate of 16.1 percentage points to 44.5 percentage points below (F=7.03). Finally, for

[37] I also conducted an effect heterogeneity tests on students identified by UB project directors as "most likely to serve" and on students by high school grade at time of application to UB. There was no evidence of a subgroup effect for either of these tests.

post-secondary completion impacts, the reference site estimate of 70.0 represented the high water mark, while the lowest impact estimate was 115.8 percentage points below that number. Put another way, at the reference site, the treatment appeared to increase post-secondary completion rates by 70.0 percentage points above the control group, while in the lowest case, UB appeared to decrease completion rates by 45.8 percentage points below the control group estimate (F=9.64). Such large variations are not especially surprising given that some sites had as few as four students. Unfortunately, there is relatively little detail about site-level program implementation with which to try to explain this variation.

Table 4.11. Tests of Effect Heterogeneity for Educational Outcomes

Effect Heterogeneity	High School Graduation	12th Grade GPA	PSE Enrollment	PSE Completion
Eligibility				
Ineligible	0.079*	0.128*	0.020	0.107*
	(0.036)	(0.062)	(0.038)	(0.046)
Eligible	0.033*	0.012	0.031*	0.023
	(0.012)	(0.036)	(0.015)	(0.025)
Difference	0.046	0.117+	-0.011	0.084+
	(0.041)	(0.070)	(0.041)	(0.050)
Algebra or above in 9th grade				
Yes	0.026	0.050	0.013	0.070**
	(0.016)	(0.044)	(0.015)	(0.025)
No	0.067*	0.061	0.020	0.050
	(0.026)	(0.048)	(0.027)	(0.038)
Difference	-0.041	-0.011	-0.007	0.020
	0.030	0.066	0.030	0.045
9th grade GPA > 2.5				
Yes	0.049**	0.069*	0.035*	0.077*
	(0.015)	(0.020)	(0.017)	(0.030)
No	0.023	0.059	0.009	0.033
	(0.026)	(0.047)	(0.021)	(0.033)
Difference	0.025	0.009	0.045	0.043
	0.001	0.002	0.001	0.001
Student expects B.A. or above				
Yes	0.037**	0.037	0.038**	0.046
	(0.012)	(0.037)	(0.013)	(0.024)
No	0.100*	0.102	-0.012	0.049
	(0.042)	(0.071)	(0.045)	(0.046)
Difference	-0.063	-0.065	0.050	-0.003
	(0.044)	(0.076)	(0.043)	(0.047)

Table notes: PSE= post-secondary education. Clustered standard errors are in parentheses. The response variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Sampling Weights

To understand the effect of weight adjustments on effect estimates I ran the Horizons, unadjusted and adjusted models using post-stratification weights, post-stratification weights trimmed at the 75th percentile and no weights (tables 4.12) (Asparouhov and Muthen, 2005). I include the estimates developed by MPR using non-response weights as a point of reference.

Those estimates represent their main published findings. Table 4.12 shows that neither trimming nor the use of no weights serves to increase the effect estimates. For the covariate-adjusted model, neither trimming nor the use of no weights serves to increase the effect estimates, and the point estimates are virtually identical under trimming and unweighted estimation. Moreover, the use of trimmed weights appears to have little influence on the point estimates. This finding suggests that trimming is a reasonable strategy (Asparouhov, and Muthen, 2005). In addition, the point estimates for the unadjusted model are virtually identical under trimming and unweighted estimation. I also note that my estimates are less sensitive to weights when covariates included in the model.

Table 4.12. Sensitivity of Effect Estimates to Weighting Assumptions- Horizons and Unadjusted Models

Model	Weights	High School Grad	Post-secondary Enrollment	Post-secondary Completion
Horizons	Horizons non-response	-0.010 n.p.	0.016 n.p.	0.130* n.p.
	Horizons post-stratification	-0.020 (0.027)	0.004 (0.026)	0.112 * (0.041)
	Trim at 75th percentile of post-stratification	0.010 (0.013)	0.022 (0.018)	0.058* (0.026)
	No weights	0.011 (0.014)	0.026 (0.017)	0.048+ (0.026)
Unadj.	Horizons post-stratification	0.006 (0.027)	0.005 (0.028)	0.003 (0.037)
	Trim at 75th percentile of post-stratification	0.034* (0.015)	0.016 (0.017)	0.029 (0.024)
	No weights	0.034* (0.013)	0.016 (0.015)	0.031 (0.023)
Covariate Adjust.	Horizons post-stratification	0.030 (0.024)	0.013 (0.025)	0.046 (0.038)
	Trim at 75th percentile of post-stratification	0.045* (0.016)	0.029+ (0.016)	0.046* (0.025)
	No weights	0.046* (0.012)	0.029+ (0.016)	0.047* (0.023)

Table Notes: Standard errors are in parentheses. 12th grade GPA was excluded from this sensitivity analysis, as it is insensitive to weighting assumptions. Horizons high school outcomes are based on 3rd survey wave responses. Horizons PSE outcomes are based on 5th survey wave responses and do not include SFA or NSC data. Unadj. = Unadjusted. Covariate Adjust. =Covariate Adjustment .The unadjusted and covariate variables use student responses from survey waves 2-4 for high school outcomes and survey waves 3-5 for post-secondary outcomes
+ $p < .10$, * $p < .05$, ** $p < .01$, *** $p < .001$

Summary of Major Findings

I started this chapter by demonstrating that I could approximately replicate the effect and impact estimates produced by MPR researchers and COE. Subsequently, I produced my effect estimates for the studied outcomes. I found evidence of a positive causal effect of UB on high school graduation and post-secondary completion, and limited evidence of a positive causal relationship between UB and post-secondary enrollment. In addition I investigated whether certain subgroups of students are affected differentially by the treatment. I found limited evidence to infer that disadvantaged students who might typically be declared ineligible for UB participation based on past behaviors or educational expectations, experience gains in rates of post-secondary completion that exceeds typically eligible students. I explored the question of whether treatment impact estimates varied significantly across sites and found preliminary evidence to suggest that there was non-random variation of the treatment impacts.

One finding, which on its face might appear contradictory, is that UB appeared to be more effective in promoting post-secondary completion than post-secondary enrollment, as shown in table 4.5. One plausible, albeit speculative explanation for this apparent discrepancy is that post-secondary enrollment was a relatively easy goal to meet since it might have required as simple a step as signing up for a course at a community or junior college. Admission to a degree program or even course attendance is not strictly required to meet the definition of enrollment. However, completing a course of study and receiving a certificate or degree is much more difficult, hence the added value of UB is more readily seen.

My results differ from those presented in the literature. MPR researchers reported that UB had no effect on high school graduation (Myers, et al., 2004). Neither MPR researchers nor

COE found evidence of effect heterogeneity for any studied subgroup (Cahalan, 2009; Seftor, et al., 2009). The sources of these differences can be traced to disparities in analysis sample construction, different approaches used in the creation of the estimation equations, and varying strategies used to address questions of internal and external validity.

I was able to replicate or approximate prior published results, which suggested that I had access to the same data sets and that I was able to employ the same or very similar statistical methods and data assumptions as prior researchers. This extended process of replication allowed me to uncover methodological issues in MPR researchers prior designs and analyses, which once addressed formed the basis for my different effect estimates.

I identified and addressed two potential threats to internal validity. Previously excluded observations were reincorporated into the data, giving a more accurate picture of the educational outcomes of Horizons students. Treatment and control imbalances, with biases that favored the control group were rectified through the use of covariate adjustments. In addition I identified and addressed two potential threats that primarily threaten external validity. Problems that threatened external validity were: 1) the use of a sample selection process, which bypassed a number of eligibility screening criteria normally employed by UB sites, and therefore generated a randomized control trial sample that is different from the typical set of program eligible students and, 2) the use of arguably incorrect probability weighting schemes.[38]

Once I corrected the problems I was able to produce what I consider to be more accurate estimates of the treatment effect. I described the reasons why my preferred estimates would have

---

[38] In addition I conducted a series of analyses on other less important outcomes, which I mention here for the sake of completeness. I did not find evidence of an effect of treatment on educational expectations, however this test was hampered by lack of analyzable data. I also did not find evidence of an effect of treatment on the type of institution to which the students applied.

higher internal and external validity in chapter 3. I subjected these estimates to sensitivity analyses in order to test the robustness of my results. Having established the credibility of my results I now turn to a discussion of what it means.

Chapter 5: Discussion

Over the past several decades a number of efforts have led to improvements in the educational equity of traditionally disadvantaged students. Federal legislation efforts such as the GI Bill and school desegregation, led to improved access to better education for lower performing students (Harris and Herrington, 2006; Gamoran, 2007). Strategies advocated by the National Commission on Excellence in Education (i.e., "A Nation at Risk") indicate a small shift in emphasis from lower performers to moderate performing students, while still advocating educational equity (Harris and Herrington, 2006). These efforts preceded the rise of government-based accountability programs, which started in the 1990's. Notably, the National Goals Education Panel (1990) and the Goals 2000 Act (1994) advanced the use of educational standards to bring about reduced levels of educational inequality (Gamoran, 2007). No Child Left Behind (NCLB) represents the most recent standards-based reform policy. The goal of NCLB is to have 100% of all students disadvantaged or not, achieve proficiency or higher on state standardized tests.

Given the potential severity of sanctions under NCLB, which at its most extreme result in school closure, it is perhaps not surprising that some schools adopt strategies to artificially raise scores on high-stakes tests (Loeb and Figlio, 2011). Strategies include classifying low-performing students as special education, and manipulating the test-taking population to exclude low-performing students (Loeb and Figlio, 2011). It is also not surprising that some schools appear to concentrate resources on those students who are at the margin of raising their test scores to the proficient level (Loeb and Figlio, 2011).

It seems unlikely that these strategies could improve long-term outcomes, since even more substantive efforts to improve student learning, such as increased time on task, have no

clear link to health or earnings. Arguably then, these strategies are also likely to do little to improve the long-term economic and health-related outcomes of disadvantaged students, especially those who are the lowest performers. Put another way, schools may adopt strategies that show improvement as measured by NCLB's criteria, (i.e., targeting resources at students who score just below the proficiency cut point), without reducing gaps in educational achievement, income levels or health status.

In contrast UB targets resources at disadvantaged students to improve their high school graduation rates, post-secondary enrollment rates and post-secondary completion rates. These improved educational outcomes are in turn, important for improving the life successes of disadvantaged students in three ways. First, higher levels of education increase worker productivity by increasing their stock of cognitive and non-cognitive skills (Becker, 1962; Mincer, 1974; Cameron and Heckman, 1993; Heckman and Rubinstein, 2001; Heckman, Stixrud, and Urzua, 2006). Second, these more productive workers are rewarded with higher lifetime earnings (Mincer, 1974; Cameron and Heckman, 1993; Kane and Rouse, 1995; Wolfe and Haveman, 2002). Third, an increase in the number of years of schooling is associated with better life outcomes including better health, lower rates of incarceration, and longer life expectancy (Wolfe and Haveman, 2002; Heckman, Stixrud, and Urzua, 2006).

UB is an effective means of increasing high school graduation rates. To the best of my knowledge, this is the first paper that finds causal evidence of this effect.[39] MPR researchers looked directly at the effects of UB, and did not find evidence of this effect (Myers and Moore, 1997; Myers and Schirm, 1999; Myers, et al., 2004). However, Lavy found causal evidence to

[39] Other researchers have reported that Upward Bound increases high school graduation rates but these findings are subject to selection effects. For example see Walsh, 2011.

suggest that disadvantaged students in earlier grades realized improved educational outcomes when exposed to a suite of intervention components, many of which are found in UB (Lavy, 2010; Lavy 2012). Specifically, he found that increasing the overall instructional time, the instructional time spent on core subjects, and the amount of homework assigned had a positive effect on the test scores. The components mentioned by Lavy align with the following UB program elements: increased instructional time delivered via after school courses, academic tutoring, after school coursework focused on core courses and group work assignments that students worked on during class time.

UB is also effective at increasing the post-secondary graduation rates of first-generation and impoverished students. However, it is potentially even more effective in raising the post-secondary graduation rates of students who might have typically been ineligible for UB in non-experiment years (i.e., students with behavioral or academic problems, or students with low educational expectations. This finding is noteworthy because 1) it is new, and 2) because it suggests that using measures of merit to determine if a student is eligible for an aid program might exclude from the applicant pool the very students that the program has, in theory, identified as the target group for intervention.

One possible lesson to be drawn from NCLB is that policies that cause resources to be allocated to students who are near a threshold (i.e., proficient), might not be as effective in improving the educational, economic and social outcomes of the more disadvantaged students, as programs that provide all students with more rigorous content and better learning opportunities (Harris and Herrington, 2006). Disadvantaged students are often quite different from low-performing students. For example, approximately 47 percent of the disadvantaged students in the

UB experiment had 9th grade GPA's over 2.5 at baseline and about 63 percent took Algebra or above in the 9th grade (Seftor, et al., 2009). Disadvantage is correlated with low performance, but there are many high performing disadvantaged students and many low performing advantaged students.

Coleman (1988) demonstrated that social capital is important for the creation of human capital. This may mean that students with low stocks of social capital could benefit from interventions like UB, which provide the opportunity to build social capital. Interventions that aim to improve social capital may be more effective than policies aimed directly at increasing academic achievement as measured by test scores. For many students, NCLB-like interventions will be ineffective because they already have the social capital they need, but their schools do not respond to the policies by increasing academic rigor. For other students, the need for intervention is great, but the academic focus dictated by NCLB-like interventions is too narrow in scope, ignoring the social capital that may be the greater problem.

Limitations

The findings I have presented in this paper are subject to a number of limitations. The effect of UB on post-secondary outcomes appear to be sensitive to the choice of covariate adjustment models, but the baseline imbalances suggest that failing to adjust for covariate imbalance would likely bias the estimates downward. The pattern of results is consistent with this interpretation.

My bounds analysis shows that if outcome data were missing because treatment assignment had no effect on educational outcomes while assignment to control had a favorable effect, then the treatment effect I estimated becomes insignificant for all outcomes.

The existing weighting structures affected my ability to clearly establish external validity. Insufficient data and documentation were available to reproduce Horizon study weights. I was able to develop a set of trimmed weights, which had lower mean squared errors and could be considered demonstrably better than the weights used in the Horizons study. The used of trimmed weights allowed me to establish my findings as externally valid.

Also, if either of Horizon study post-stratification or non-response weights were in fact correct, my effect estimates would be insignificant. If I had settled on using the Horizon study non-response weights, between 13% and 19% of observations I consider valid would have been excluded, since this weighting structure assigns a weight of zero to observations that were not collected during the 3rd follow up survey for high school outcomes and the 5th follow up survey for post-secondary outcomes, and it assigns a weight of one to responses collected from SFA and NSC datasets.

As a rejoinder to those who would argue for using Horizon study weights, I will summarize the evidence that the Horizon study weights are subject to limitations. Strict adherence to Horizon study stratification algorithms arguably places project 69 in a different stratum than the one it was assigned to, and reduces its sample weight from over 26% to fewer than 3% (Cahalan, 2009). Post-stratification weights are based upon the distributions of student characteristics for the UB population. This population was pre-screened to eliminate undesirable student behavior patterns and educational expectations, but the post-stratification weights ignore the fact that many students were ineligible. Horizons non-response weights also ignore between

13% and 19% of valid student responses. This analysis highlights the importance of deriving the correct weights, if weights are to be used at all.

In addition, the screening tool I built to identify these students, while useful, is perhaps somewhat inaccurate. Applying it would have screened out 34.6% of applicants, versus the 43.6% actually screened out. One reason for the difference in percentages of screened out students may be measurement error in my gauge of ineligibility. This difference leaves room for substantial improvement in the development of a measurement tool, and suggests the possibility of further testing through additional analysis using the retrospective data from the Horizons study.

There were other specific limitations as well. Self-reported survey data are subject to error including over-reporting and underreporting of events (Groves, 2009). Nonetheless other researchers do use self-reported survey data to answer important education research questions (for example see Adelman, 2006). This use of self-reported data includes the use of US Census data to answer education research questions (for examples see Day and Baum, 2000; Fry, 2011, Fry and Lopez, 2012).

To address under-reporting of post-secondary enrollment and completion, MPR researchers used supplemental SFA and NSC data. [40] These supplemental data sources were also subject to limitations. MPR researchers noted that the key limitation in using Pell Grant data was

[40] If the NSC or SFA data sets showed evidence of post-secondary enrollment, the student self-reported response was recoded accordingly whether the initial response was missing, or the student indicated that he or she had not enrolled in and/or completed post-secondary education.

measurement error because receipt or non-receipt of the grant did not equal enrollment or non-enrollment (Seftor et al., 2009). [41]

Also COE (2012) noted that limitations in using early NSC data were under coverage, and bias. Specifically, COE noted that:

> Against PPSS's recommendation and that of the IES external reviewers to be "conservative in use of NSC" Mathematica chose to report in the text tables and conclusions only those estimates that use NSC data for non-responders to the Fifth Follow-up—coding the 25 percent of the sample who were survey non-responders and who were not found in NSC as "not having a degree or certificate." (p. 29)
>
> As originally recommended by PPSS Technical Monitors, Mathematica should not use NSC data from this early period for enrollment estimates as NSC does not have enough coverage (25 percent) and there is evidence of bias due to clustering of UB participants in the grantee institutions. (p. 30)

There is reason to think that degree coverage by NSC was not as complete for 2-year schools as it was for four-year schools. Table C.7 of the 5th follow up report details the effect of UB on degree completion by type of degree granting institution. The ratio of NSC degrees reported to survey degrees reported by type of institution for control students only was 0.53 for four-year colleges, while the 2-year school ratio was 0.41, and the credentialing school ratio was 0.07. These declining ratios suggest a relative under coverage for the 2-year and credentialing institutions.

In addition, National Student Clearinghouse Data indicates that while enrollment coverage for all institutions rose from 86.6% to 92.1% during Fall 2003-Fall 2009 (a period corresponding to the Horizons final report), coverage at two-year private schools did not keep pace. Coverage at not-for-profit schools dropped from 43% in Fall 2003 to 36% by Fall 2009,

[41] To reiterate, the lack of access to NSC or complete SFA data impacted my ability to replicate Horizons study post-secondary findings in general. Furthermore, I was not able to reliably test how sensitive my post-secondary results were to the inclusion of these data.

while coverage at for-profit schools rose from 11.8% in Fall 2003 to 14.2% by Fall 2009 (National Student Clearinghouse Research Center, 2012). Put another way in 2003 the NSC database picked up about 37,000 enrollees at two-year private schools while IPEDS puts the number at 285,000. For 2009, the corresponding numbers are 57,000 from NSC and 420,000 from IPEDS (National Student Clearinghouse Research Center, 2012).

There are also more general limitations to randomized trials that are applicable. For example treatment students may have benefited from the "Hawthorne effect" while control students may have become resentful or demoralized as a result of not being chosen, which would exaggerate the treatment effect (Shadish, Cook, and Campbell, 2002). Also, treatment students who were "no-shows" might have decided not to enroll because they thought the program might not be effective for them, which would constitute selection bias.

Appropriate care should be taken in extrapolating from my findings. The entering UB class of 2013 may bear little resemblance to those who took part in the Horizons study. While precise comparisons are difficult to make, recent program participants from 2000-2001 appear to closely resemble program participants from the 1990's on a limited set of observable demographics, which should mitigate a portion of the extrapolation risk (Moore, et al., 1997; Cahalan, 2004). Also the Horizons study design parameters purposely excluded any project site that had been in existence for less than three years, was administered by a community organization or that had less than two applicants for each program slot, and so my research findings may not be applicable to those sites. Finally students who participated in the study were motivated to apply and in many instances, motivated to attend UB course sessions, and this level of motivation is not certain to exist in students in general. In fact an ad hoc comparison between

Horizons students and students who attended UB feeder schools suggested that Horizons student have higher educational expectations, even though their parents appear to have less formal education (table A.1).

A comparison of the program of instruction between the 1990's and 2000-2001 does suggest that the academic offering has concentrated on presenting more math and science courses year round, and more advanced math and sciences courses during the school year, as well as foreign language, while possibly de-emphasizing fine arts and computer software and applications. These differences in the composition of the intervention also suggest caution in extrapolating from then to now (Moore, et al., 1997; Cahalan, 2004).

Future Research

One of the strong original motivating reasons to pilot and then launch a nationwide UB program was the belief that increased levels of education would translate into better employment opportunities and better wages. A logical extension to the Horizons study body of work would be a follow-up study that examines if UB caused higher rates of employment and higher incomes.

A second new research opportunity would explore the effects of different recruiting practices (i.e., targeting disadvantaged students). An investigation of this type could be performed at UB sites where oversubscription occurs naturally. Current recruiting practices may contribute to suboptimal post-secondary outcomes. UB program directors target students who show college potential and have low frequencies of behavioral and academic problems. This targeting strategy bypasses students who are or who might become motivated to earn a degree or certificate, but who have a history of behavioral or academic problems, or who are unsure about their future academic goals.

Increasing educational expectations and by extension, motivation might be as simple as providing disadvantaged students with information about the value of post-secondary education. A recent experiment conducted by Oreopoulos and Dunn (2012) found that educational expectations were raised for disadvantaged high school students in Toronto, Canada who were exposed to a three-minute video about the benefits of a post-secondary education and were invited to use a financial aid calculator to estimate financial support. Effects were larger for those students who were unclear about their future academic goals.

Another reason to broaden the footprint of UB is that barely 20% of socioeconomically eligible students have access to a UB project site (Moore et al., 1997). Given the high per student costs of UB, the high percentage of socioeconomically eligible students who do not have access to a UB program site, and given the rise in on-line education, the time seems right to launch a pilot program which would gauge the effectiveness of a hybrid UB offering, combining an on-line learning curriculum with an in-person campus experience. This could serve as a cost-effective way to deliver similar increases in outcomes.

Cost effectiveness is an important consideration especially in light of the declining productivity curve for degree production Harris and Goldrick-Rab (2010). While I find UB to be an effective intervention for increasing post-secondary graduation rates, it is not clear if it is economically efficient, and this question should be more thoroughly investigated for the Horizons study students.

As an example, Harris and Goldrick-Rab (2010) found that in order to equate the cost effectiveness of UB with less expensive alternatives, a class of 100 UB students would need to earn an additional 18 degrees. Using my effect estimates, I calculated that UB produced

approximately 4.7 additional degrees per 100 students. Using my effect heterogeneity estimates, I calculated that an incremental 8.6 degrees would be produced per 100 typically ineligible students. While apparently short of the break-even point, these crude calculations likely understate the additional degrees produced because my denominator includes students who did not graduate from high school and did not enroll in a post-secondary institution.

Another prospective line of research would investigate the effects of a version of UB aimed solely at increasing high school graduation rates. I argue that UB as currently constituted is not optimized to increase high school graduation rates and does not target students who would optimally benefit from such a program. For example, the applicant screening process used by many UB program directors actively excludes students who do not indicate that they are going on to post-secondary school, and this hurdle might discourage student who would graduate from high school as a result of UB.

More generally, it is difficult to optimize a program to achieve the greatest effect on any single outcome when an intervention has multiple intended outcomes, which is the case with UB. Time spent on activities such as learning how to apply for financial aid and learning what college life is like is time not spent on high school mathematics or reading. While activities like filling out the Free Application for Federal Student Aid (FAFSA) forms may have an indirect effect on high school graduation it is likely that activities that bear directly on meeting high school requirements or increasing student engagement in school will yield greater benefits.

Extending this idea, a further line of research could investigate the effects of versions of UB focused on increasing post-secondary enrollment and completion rates by differing types of institution. It is possible that all of the post-secondary treatment benefits to date accrued to

students who pursued a four-year degree. However, as Kane and Rouse (1995) have suggested, economic benefits also accrue to students who complete two-year degrees and complete individual years of college without earning a degree.

The potential also exists to investigate program components and student characteristics that are associated with impact variations across sites. I explored this question and found some evidence to suggest that impact variation exists. It is possible this question could be reframed as a causal mediation analysis, which may represent a contribution to an emerging field of study in education policy. Because of the Congressional prohibitions against future RCT's of UB, this question would need to be investigated using a quasi-experimental design.

Educational achievement and attainment are critical to economic, social and health outcomes. A number of initiatives designed to broaden the pathways to college have been introduced in the last few decades. Some of those initiatives have moderately broadened the educational pathways or possibly improved the educational outcomes of the target populations (Kleiner and Lewis, 2005; Waits, Setzer and Lewis, 2005; Domina, 2009). For example, in a recent experiment, students from low and moderate-income families who were randomly assigned to receive hands-on assistance in completing financial aid application forms as well as general information about financial aid were more likely to turn in the application, enroll in college in the ensuing year, and receive higher levels of financial aid than the control group who received only general information about financial aid (Bettinger, Long, Oreopoulos, and Sanbonmatsu, 2009).

Rich-poor or minority-majority education gaps remain a problem. The rate at which these gaps are narrowed has apparently slowed to near zero in the last 15 years. The data I have

presented here suggests that a partial solution to the problem may already exist in Upward Bound.

References

Adelman, C. 1999. Answers in the Tool Box: Academic Intensity, Attendance Patterns, and Bachelor's Degree Attainment. Washington, DC: U.S. Department of Education.

Adelman, C. (2006). The toolbox revisited: paths to degree completion from high school through college. Washington, D. C.: Office of Vocational and Adult Education, U.S. Dept. of Education.

Asparouhov, T. and Muthen, B. (2005). Testing for informative weights and weights trimming In multivariate modeling with survey data. Section on Survey Research Methods. 3394-3399.

Albee, Amy (2005). A Cost Analysis of Upward Bound And GEAR UP. Unpublished manuscript. Tallahassee, FL: Florida State University.

Alexander, K. L., Entwisle, D. R., and Horsey, C. S. (1997). From First Grade Forward: Early Foundations of High School Dropout. Sociology of Education, 70(2), pp. 87-107.

Allison, Paul D. (2009) "Missing data." Pp. 72-89 in The SAGE Handbook of Quantitative Methods in Psychology, edited by Roger E. Millsap and Alberto Maydeu-Olivares. Thousand Oaks, CA: Sage Publications Inc.

Allison, Paul D. (2002) Missing Data. Thousand Oaks, CA: Sage Publications

Aud, S., Hussar, W., Kena, G., Bianco, K., Frohlich, L., Kemp, J., Tahan, K. (2011). The Condition of Education 2011 (NCES 2011-033). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

Bailey, Martha J., and Dynarski , Susan M. (2011). Gains and Gaps: Changing Inequality in U.S. College Entry and Completion. NBER Working Paper No. 17633. December 2011

Becker, Gary S. (1962) Investment in Human Capital: A Theoretical Analysis Journal of Political Economy Vol. 70, No. 5, Part 2: Investment in Human Beings (Oct., 1962), pp. 9-49. The University of Chicago Press.

Becker, S. (2012). An Intersectional Analysis of Differential Opportunity Structures for Organized Crime Prevention Efforts. Race and Justice.

Bettinger, E., Long, B., Oreopoulos, P., and Sanbonmatsu, L. (2009). The role of Simplification and Information in College Decisions: Results from the HandR Block FAFSA experiment (NBER Working Paper No. 15361). Cambridge, MA: National Bureau of Economic Research.

Bloom, H. (2012, March). Impact Variation: How Do You Know It When You See It? Speech presented at SREE, Washington, DC.

Bloom, H. (Ed.) (2005). Learning More From Social Experiments: Evolving analytical approaches. New York: Russell Sage Foundation.

Bloom, H. (1984) Accounting for No Shows in Experimental Evaluation Designs. Evaluation Review, vol. 8, 1984.

Bloom, H. and Michalopoulos, C. (2010). When is the Story in the Subgroups? Strategies for Interpreting and Reporting Intervention Effects for Subgroups. MDRCWorking Paper. November. http://www.mdrc.org/publications/551/full.pdf

Bourdieu, P. 1986. 'The Forms of Capital.' Pp. 241-58 in Handbook of theory and research for the sociology of education, edited by John G Richardson. New York: Greenwood Press

Burkheimer, G. J., Levinsohn, Koo and French. (1976). Evaluation Study of the Upward Bound Program: Volume IV. Final report

Cahalan, M. W., Curtin, T. R. (2004). Research Triangle Institute, and Office of Postsecondary Education (ED), Washington, DC. A Profile of the Upward Bound Program: 2000-2001. US Department of Education.

Cahalan, M. W. (2009). Addressing Study Error in the Random Assignment National Evaluation of Upward Bound: Do the Conclusions Change? Council for Opportunity in Education, USDOE.

Cameron, S. V., and Heckman, J. J. (1993). The Nonequivalence of High School Equivalents. Journal of Labor Economics, 11, 1–47

Chowdhury, S., Khare, M., and Wolter, K. (2007), "Weight Trimming in the National Immunization Survey," Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association.

Cloward, Richard A. and Ohlin, Lloyd E. (1960). Delinquency and Opportunity; a Theory of Delinquent Gangs. Glencoe, Ill., Free Press [1960]

COE (2012). Request for Correction for the Report: The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation (Referred to as the Mathematica Fifth Follow Up Report), Prepared by Mathematica Policy Research. Council for Opportunity in Education

Cohen, Jacob (1992), "A Power Primer", Psychological Bulletin 112 (1): 155–159

Coleman, J. S. (1988). Social capital in the creation of human capital. The American Journal of Sociology [AJS], 94 (Supplement), S95-S120.

Davis, J. A., and Kenyon, C. A. (1976). A Study of The National Upward Bound and Talent Search Programs. Final Report. Volume I.

Day, Jennifer Cheeseman, and Kurt J. Bauman. 2000. "Have We Reached the Top? Educational Attainment Projections of the U.S. Population." Population Division Working Paper No. 43. Washington, DC: U.S. Census Bureau, May.

Domina, T. (2009). What Works in College Outreach: Assessing Targeted and Schoolwide Interventions for Disadvantaged Students, Educational Evaluation and Policy Analysis, 31(2), 127-152.

Elliott, M.R. (2008), "Model Averaging Methods for Weight Trimming," Journal of Official Statistics, 24, 517-540.

Elwert, F. and Winship, C. (2010) "Effect Heterogeneity and Bias in Main-Effects-Only Regression Models" with Felix Elwert, in press in Heuristics, Probability and Causality: A Tribute to Judea Pearl, Rina Dechter, Hector Geffner, and Joseph Y. Halpern (eds.).

Esbensen, F., Peterson, D., Taylor, T. J., and Freng, A. (2009). Similarities and Differences in Risk Factors for Violent Offending and Gang Membership. Australian and New Zealand Journal Of Criminology (Australian Academic Press), 42(3), 310-335. doi:10.1375/acri.42.3.310

Faul, F., Erdfelder, E., Lang, A.-G., and Buchner, A. (2007). G*Power 3: A Flexible Statistical Power Analysis Program for the Social, Behavioral, and Biomedical Sciences. Behavior Research Methods, 39, 175-191.

Field, K. (2007). Are the Right Students 'Upward Bound?'. Chronicle of Higher Education, Vol. 53 Issue 50, p16-16 Retrieved from EBSCOhost on 3/01/2010.

Freedman, D.A. (2008). On Regression Adjustments to Experimental Data. Advances in Applied Mathematics vol. 40 (2008) pp. 180–93.

Fry, Richard. (2011) Hispanic College Enrollment Spikes, Narrowing Gaps with Other Groups. Pew Hispanic Center. Retrieved from http://www.pewhispanic.org on 3/13/2012

Fry, Richard and Lopez, Mark. (2012). "Hispanic Student Enrollments Reach New Highs in 2011." August. Washington, DC: Pew Hispanic Center.

Fultz, M. and Brown, A. L. (2008). Historical perspectives of African American males as subjects of education policy. American Behavioral Scientist, 51 (7), 854-871.

Gamoran, Adam, Editor. (2007). Standards-Based Reform and the Poverty Gap: Lessons for No Child Left Behind. Washington, DC: Brookings Institution Press.

Gándara, P., and Bial, D. (2001). Paving the Way to Postsecondary Education: K-12 Interventions for Underrepresented Youth. Washington, D.C. 2001. National Center for Education Statistics.

Gelman, Andrew, and Jennifer Hill. (2007). Data Analysis Using Regression and Multilevel/Hierarchical Models. Cambridge: Cambridge University Press.

Greenleigh , Arthur. (1970) "History of Upward Bound," in Upward Bound 1965-1969: A History and Synthesis of Data on the Program in the Office of Economic Opportunity

Groutt, John. (2011). A Role Foundations Played in Advancing Opportunities in Higher Education for American Poor and Minorities. GIVING: Thematic Issues on Philanthropy and Social Innovation, Social Justice Philanthropy, new series, no. 2 (2011): 39-54. Bononia University Press. Bologna, Italy

Groutt, John, and Hill, Calvin (2001). Upward Bound: In the Beginning. Opportunity Outlook. April, 2001, pages 26-33.

Groves, R. M. (2009). Survey methodology. 2nd ed. Hoboken, N.J.: Wiley.

Gullatt, Y., and Jan, W. (2003). How do Pre-collegiate Outreach Programs Impact College Going Among Underrepresented Students? The Pathways to College Network Clearinghouse.

Hanushek, Eric A. and Lindseth, Alfred A. (2009) Schoolhouses, Courthouses, and Statehouses: Solving the Funding-Achievement Puzzle in America's Public Schools (Princeton, NJ: Princeton University Press, 2009).

Hanushek, Eric A. and Woessmann, Ludger. (2010). The Economics of International Differences in Educational Achievement (NBER Working Paper No. 15949). Cambridge, MA: National Bureau of Economic Research.

Harris, D. N., and Herrington, C. D. (2006). Accountability, Standards, and the Growing Achievement Gap: Lessons From the Past Half-century. American Journal of Education, 112(2), 209-238.

Harris, D. N. (2012). Improving the Productivity of American Higher Education through Cost-Effectiveness Analysis. Wisconsin Center for the Advancement of Postsecondary Education (WISCAPE)

Harris, D.N., and Goldrick-Rab, S. (Forthcoming) Improving the Productivity of Education Experiments: Lessons from a Randomized Study of Need-Based Financial Aid. Education Finance and Policy

Harris, D. and Goldrick-Rab, S. (2010) The (un)productivity of American higher education: A template for cost-effectiveness analysis to help overcome the "cost disease". Wisconsin Center for the Advancement of Postsecondary Education (WISCAPE)

Heckman, J., Stixrud, J., & Urzua, S. (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. Journal of Labor Economics, 24(3), 411-482. Retrieved April 10, 2009, from Business Source Elite database.

Heckman, J. J., and Y. Rubinstein. The Importance of Noncognitive Skills: Lessons from the GED Testing Program. American Economic Review, 91(2), 2001, 145-9

Heckman, J.J. and LaFontaine, Paul A. (2010). "The American High School Graduation Rate: Trends and Levels," The Review of Economics and Statistics, MIT Press, vol. 92(2), pages 244-262, 01.

Henry, K., and Valliant, R. (2012) Methods for Adjusting Survey Weights When Estimating a Total. Retrieved February 20, 2013 from fcsm.gov

Hirschfield, Paul, and Gasper, Joseph. (2011). The Relationship between school engagement and delinquency in late childhood and early adolescence. Journal of Youth and Adolescence, 40(1), 3-22.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. Political Analysis, 15(3), 199.

Holland, P. W. (1986). Statistics and Causal Inference. Journal of the American Statistical Association, 81(396), pp. 945-960.

Iacus, SM, King G, Porro G. (2011). Causal Inference Without Balance Checking: Coarsened Exact Matching. Political Analysis.

Imai, Kosuke and Marc Ratkovic. (2010). Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation." Political Analysis (2011) 19:1–19

Imai, K., King, G., and Stuart, E. A. (2008). Misunderstandings Between Experimentalists and Observationalists About Causal Inference. Journal of the Royal Statistical Society. Series A (Statistics in Society), 171(2), pp. 481-502.

Imbens, G. and J. Angrist (1994). Identification and Estimation of Local Average Treatment Effects. Econometrica 62, 467–476.

James, Donna Walker, Sonia Jurich and Steve Estes (2001). Raising Minority Academic Achievement: A Compendium of Education Programs and Practices. Washington, DC: American Youth Policy Forum.

Jencks, C., and Phillips, M. (1998). The Black-White Test Score Gap. Washington, D.C.: Brookings Institution Press.

Jones, Michael P. (1996). Indicator and Stratification Methods for Missing Explanatory Variables in Multiple Linear Regression. Journal of the American Statistical AssociationVol. 91, No. 433 (Mar., 1996), pp. 222-230

Kane, T., and. C. Rouse. (1995): Labor market Returns to Two- and Four-Year College. American Economic Review, 85(3)

Kleiner, B., and Lewis, L. (2005). Dual Enrollment of High School Students at Postsecondary Institutions: 2002–03 (NCES 2005–008). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Kreager, Derek A., Kelly Rulison, and James Moody. 2011. "Delinquency and the Structure of Adolescent Peer Groups." Criminology 49(1):95-127.

Lavy, Victor (2012). Expanding School Resources and Increasing Time on Task: Effects of a Policy Experiment in Israel on Student Academic Achievement and Behavior. (NBER Working Paper No. 18369). Cambridge, MA: National Bureau of Economic Research.

Lavy, Victor (2010). Do Differences in Schools' Instruction Time Explain International Achievement Gaps? Evidence from Developed and Developing Countries. (NBER Working Paper No. 16227). Cambridge, MA: National Bureau of Economic Research.

Lee, D. R. and Cohen, J. W. (2008) Examining strain in a school context. Youth Violence and Juvenile Justice: An Interdisciplinary Journal, 6, 2, 115-135.

Liu, B., Ferraro, D., Wilson, E., and Brick, M. (2004). Trimming Extreme Weights in Household Surveys. ASA Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association, 2004, 3905- 3911.

Loeb, S., and Figlio, D. (2011). School accountability. In E. A. Hanushek, S. Machin, and L. Woessmann (Eds.), Handbook of the Economics of Education, Vol. 3 (pp.383-423). San Diego, CA: North Holland.

Magnuson, Katherine and Jane Waldfogel, Editors. (2011). Steady gains and Stalled Progress: Inequality and the Black-White Test Score Gap. New York: Russell Sage Foundation, 2011.

Manski, C. F. (1995). Identification Problems in the Social Sciences. Cambridge, Mass.: Harvard University Press.

Manski, C. F. (2003). Partial Identification of Probability Distributions. New York: Springer.

Merton, R. K. (1938). Social structure and anomie. American sociological review, 3(5), 672-682.

Mincer, Jacob. (1974). Schooling, Experience and Earnings New York: National Bureau of Economic Research.

Moore, M. T., Fasciano, N. J., Jacobson, J. E., Myers, D., and Waldman, Z. (1997). The National Evaluation of Upward Bound. A 1990's View of Upward Bound: Programs Offered, Students Served, and Operational Issues. Background Reports: Grantee Survey Report, Target School Report.

Morgan, Stephen L. and Winship, Christopher. (2007). Counterfactuals and Causal Inference: Methods and Principles for Social Research. Cambridge: Cambridge University Press.

Myers, D. E., and Moore, M. T. (1997). The National Evaluation of Upward Bound. Summary of First-year Impacts and Program Operations. Executive Summary Journal of Educational Opportunity, 16(2), 61-68.

Myers, D. E., and Schirm, A. L. (1997). The National Evaluation of Upward Bound. The Short-Term Impact of Upward Bound: An Interim Report . US Department of Education. Washington, DC

Myers, D., and Schirm, A. (1999). The Impacts of Upward Bound: Final Report for Phase I of the National Evaluation. US Department of Education. Washington, DC

Myers, D., Olsen, R., Seftor, N., Young, J., and Tuttle, C. (2004). The Impacts of Regular Upward Bound: Results From the Third Follow-up Data Collection. Doc. #2004-13ED

National Student Clearinghouse Research Center. (2012). NSC Enrollment Coverage [Data file]. Retrieved from http://research.studentclearinghouse.org/working_with_my_data.php

Oreopoulos, P. and Dunn, R. (2012). Information and College Access: Evidence from a Randomized Field Experiment, NBER Working Papers 18551, National Bureau of Economic Research, Inc.

Pedlow, S., Porras, J., O.Muircheartaigh, C., and Shin, H. (2003), "Outlier Weight Adjustment in Reach 2010," Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association, 3228-3233.

Perna, L. W. (2002). Pre-college outreach programs: Characteristics of programs serving historically underrepresented groups of students. Journal of College Student Development, 43, 64-83.

Potter, F. (1988), "Survey of Procedures to Control Extreme Sampling Weights," in Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 453-458.

Potter, F. (1990). A study of Procedures to Identify and Trim Extreme Sampling Weights, Proceedings of the Section on Survey Research Methods, American Statistical Association, pp. 225-230.

PPSS (2009). Policy and Program Studies Service Report Highlight. The Impacts of Regular Upward Bound on Postsecondary Outcomes 7-9 Years After Scheduled High School Graduation: Final Report. Retrieved from http://www2.ed.gov/about/offices/list/opepd/ppss/reports.html#higher March, 2012

PPSS (2011). National Evaluation of Upward Bound [Data files and code books]. US Department of Education. Washington, DC

Reardon, Sean F. (2011) . In R. Murnane and G. Duncan (Eds.), Whither Opportunity? Rising Inequality and the Uncertain Life Chances of Low-Income Children, New York: Russell Sage Foundation Press. 2011.

The Rockefeller Foundation Annual Report for 1963 (1963). The Rockefeller Foundation 111West 50th street, NY, NY 10020. Downloaded from http://www.rockefellerfoundation.org on 2/17/2013

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. Journal of Educational Psychology 66 688–701.

Rumberger, R. W., and Larson, K. A. (1998). Student Mobility and the Increased Risk of High School Dropout. American Journal of Education, 107(1), pp. 1-35.

Rumberger, R., and Lim, S. (2008). Why Students Drop Out of School: A Review of 25 Years of Research. California Dropout Research Project. Gevirtz Graduate School of Education, UC Santa Barbara, Santa Barbara, CA

Seftor, N. S., Mamun, A., and Schirm, A. (2009). The Impacts of Regular Upward Bound on Postsecondary Outcomes Seven to Nine years After Scheduled High School Graduation. final report. US Department of Education. P.O. Box 1398, Jessup, MD 20794-1398

Shadish, W., Cook, T. and Campbell, D (2002). Experimental and Quasi-Experimental Designs for Generalized Causal. Inference. Boston: Houghton Mifflin

Shaw, C. R., & McKay, H. D. (1942). Juvenile delinquency and urban areas: A study of rates of delinquents in relation to differential characteristics of local communities in American cities. Chicago: University of Chicago Press.

Skogan, W. (1989). Communities, Crime, and Neighborhood Organization. Crime and Delinquency. 1989; 35(3):437-457.

Smith, T. M., and And Others. (1995). The Condition of Education, 1995 U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC

Smith, T. M., and And Others. (1995a). The Pocket Condition of Education, 1995 U.S. Government Printing Office, Superintendent of Documents, Mail Stop: SSOP, Washington, DC

Staats, E. B. (1974). Problems of the Upward Bound Program in Preparing Disadvantaged Students for a Postsecondary Education B-164031(1), Mar 7, 1974 United States General Accounting Office.

Staff, Jeremy, and Derek A. Kreager. 2008. "Too Cool for School? Peer Status and High School Dropout." Social Forces 87(1):445-471.

Stillwell, R., Sable, J., and Plotts, C. (2011). Public School Graduates and Dropouts from the Common Core of Data: School Year 2008-09. First Look. NCES 2011-312 National Center for Education Statistics. Available from: ED Pubs. P.O. Box 1398, Jessup, MD 20794-1398. Tel: 877-433-7827; Web site: http://nces.ed.gov/.

Stuart, E.A. (2010). Matching Methods for Causal Inference: A Review and a Look Forward. Statistical Science 25(1): 1-21.

Swail, W.S. (2001). Educational Oportunity and the Role of Pre-college Outreach Programs. College Board Outreach Program Handbook. Washington, D.C.: Educational Policy Institute. Retrieved August, 2010, from http://www.educationalpolicy.org/pdf/OutreachHandbookEssays.pdf

Trimming of School Base Weights (2003). Retrieved February 19, 2013, from http://nces.ed.gov/nationsreportcard/tdw/weighting/2002_2003/weighting_2003_base_schtrim.asp

USDOE (1988). U.S. Dept. of Education, National Center for Education Statistics. NATIONAL EDUCATION LONGITUDINAL STUDY, 1988. Chicago, IL: National Opinion Research Center [producer], 1989. Ann Arbor, MI

USDOE (2010). List of 2010 Upward Bound Grantees. Retrieved from "http://www2.ed.gov/programs/trioupbound/awards.html" on 5/2/2011

USDOE (2012). Schedule of 2011 Talent Search Costs. Retrieved from http://www2.ed.gov/programs/triotalent/funding.html" on 5/25/2012

Von Hippel, P. T. (2007). Regression with Missing Y's: An Improved Strategy for Analyzing Multiply Imputed Data. Sociological Methodology, 37(1), 83-117.

Waits, T., Setzer, J.C., and Lewis, L. (2005). Dual Credit and Exam-Based Courses in U.S. Public High Schools: 2002–03 (NCES 2005–009). U.S. Department of Education. Washington, DC: National Center for Education Statistics.

Wall, Ellen, Gabriele Ferrazzi, and Frans Schryer. (1998). Getting the Goods on Social Capital. Rural Sociology 63: 300-322.

Walsh, Rachael. (2011). Helping or Hurting: Are Adolescent Intervention Programs Minimizing Racial Inequality? Education and Urban Society May 2011 43: 370-395

Weight Trimming Adjustments for the 2005 Assessment. (2005). Retrieved February 19, 2013, from http://nces.ed.gov/nationsreportcard/tdw/weighting/2004_2005/weighting_2005_trimming_adjustments.asp

Winship, Christopher and Radbill, Larry. (1994) Sampling Weights and Regression Analysis. Sociological Methods Research November 1994 vol. 23 no. 2 230-257

Wolfe B, and Haveman R. (2002). Social and non-market benefits from education in an advanced economy. Paper presented at: Education in the 21st Century: Meeting the Challenges of a Changing World; June 2002; Boston, Mass. Available at: www.bos.frb.org/economic/conf/conf47/conf47g.pdf. Accessed May 17, 2009

Appendix

Table A.1. A Comparison of Horizons Students with Students Enrolled at Upward Bound Feeder High Schools

Variable	Horizons Sample 1992-1994	Students at Upward Bound Feeder Schools - Circa 1988 (n=284)
Low-income household	0.85	N.R.
At least one parent has a B.A.	0.06	0.27
Female	0.67	0.53
White	0.28	0.37
Hispanic	0.19	0.21
Black	0.43	0.31
Other race	0.10	0.11
Student expects to earn a B.A. or above [1]	0.81	0.69

Table Notes: Students at Upward Bound Feeder School data from the National Education Longitudinal Study of 1988 (NELS:88), using listwise deletion (USDOE, 1988). NELS:88 reports data for student cohorts originally assembled and surveyed in 8th , 10th, and 12th grade (and beyond). This data structure is not directly comparable to UB data, where students can enter the program between the times they are considered rising 9th graders until they are considered 12th graders. NELS:88 lacks comparable poverty level data making it difficult to determine how many of the 284 cases would truly meet federal eligibility requirements.
N.R. = Not Reported

[1] Survey response given while student was in 10th grade. All other responses were from the 8th grade baseline survey.