INSTITUTE for RESEARCH on POVERTY
UNIVERSITY OF WISCONSIN–MADISON

IRP Discussion Paper

WISCAPE
Wisconsin Center for the Advancement of Postsecondary Education

No. 1426-14

The Upward Bound College Access Program 50 Years Later: Evidence from a National Randomized Trial

Douglas N. Harris
Associate Professor of Economics
University Endowed Chair in Public Education
Tulane University

Alan Nathan
Ryne Marksteiner
University of Wisconsin at Madison

December 9, 2014

Acknowledgements: For their useful comments, we thank Margaret Cahalan, Howard Bloom, Jill Constantine, Adam Gamoran, Rebecca Maynard, David Meyers, Rob Olsen, Barbara Schneider, Gary Solon, and Neil Seftor and session participants at MDRC and the 2013 annual meeting of the Society for Research on Education Effectiveness. We are grateful to the U.S. Department of Education's Policy and Program Studies Service (PPSS) for providing access to the data. All remaining errors are our own.

Abstract

Upward Bound (UB) was one of the original federal *Great Society* programs of the 1960s and remains, fifty years later, the single largest college access program in the country. Recently, Congress has reduced funding and considered eliminating the program because of federal budget pressures and because the first analysis of the only national randomized trial concluded that UB had "no detectable effect." In this study, we explain problems with the study design that have made this conclusion controversial and made it difficult to identify average treatment effects that are unbiased for the program population. The study design is sufficient, however, to test other important hypotheses. We show that UB has drifted away from its original intent of serving disadvantaged students and become less efficient in the process. Specifically, over time, UB has come to serve students whose parents have higher incomes and education levels. Also, program administrators deem students ineligible based on misbehavior and other factors, even though assignment to the UB treatment increases this ineligible group's probability of graduating high school and obtaining some type of college credential by 6–10 percentage points. We show that the program is cost-effective for this group and would reduce achievement gaps if it were better targeted. In short, the government should encourage sites to get back to the program roots of serving more disadvantaged students.

Keywords: College entry; Upward Bound; Federal programs; Evaluation

I. Introduction

The *Great Society* programs of the early 1960s sought to reduce poverty and equalize social and economic opportunity across racial and income groups. In addition to welfare, health care, and affirmative action laws, policymakers recognized that unequal educational opportunities were a key cause of poverty and potentially a key lever for solving it. Three federal programs, known collectively as TRIO,[1] were therefore part of the Economic Opportunity Act of 1964 and aimed to improve college outcomes for first-generation college students and those from low-income families.

Yet, 50 years later, these efforts have fallen far short of their goals. High school graduation, college enrollment and college completion rates for low-income and minority students trail other groups by a wide margin. Sixty-five percent of African-American and Hispanic students graduate high school as compared to 82 percent of white students (Heckman and LaFontaine, 2007). The gaps are even larger across income groups (Reardon, 2011). Eighty percent of students who are in the highest income quartile attend college while 29 percent of all students in the lowest income quartile do so, resulting in a college entrance gap of 51 percentage points (Bailey and Dynarski, 2011). The role of family income is even stronger when shifting from college entry to graduation: students in the highest income quartile are six times as likely to complete college compared with the lowest income quartile.[2] These findings raise concerns about both the efficiency and equity of the education system, especially the college access programs intended to address college outcomes gaps.

[1] The three original programs are: Upward Bound, Educational Talent Search and Student Support Services. There are now eight TRIO programs, including these original three, though the term TRIO is still applied.

[2] These numbers are unconditioned upon prior college entrance and focus on bachelor's degree completion within six years (Bailey & Dynarski, 2011).

Unfortunately, there remains limited evidence about the effectiveness of college access programs generally. Since the mid-1980's approximately 500 pre-college programs have sought to improve college access for disadvantaged students (Moore, 1997). One review of 200 programs aimed at raising minority student achievement (broadly defined) showed that 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. Only one of the four studies focused on post-secondary outcomes: the federally funded national experiment with Upward Bound (James et al. 2001).

Upward Bound (UB), one of the original TRIO programs, continues to be the nation's flagship college access program. Students may enroll in UB for up to four years of high school and receive a full menu of services: after-school and weekend classes, tutoring, summer school, SAT/ACT exam preparation, mentoring, and counseling. This menu has come to define what college access programs do.

Early in its history, UB was judged to be "an incredible success story" as 70 percent of UB high school graduates enrolled in college compared with 50 percent nationally at the time of the study (Greenleigh, 1970, p.2). This descriptive study was reinforced by a 1973 matched sample study that also found much greater college attendance for UB students than for comparison students (Burkheimer, et al., 1976; Moore, Fasciano, Jacobson, Myers, and Waldman, 1997). As a result, Upward Bound (UB) was considered an exemplary pre-college program (Gandara and Bial, 2001; Gullat and Jan, 2003; James, et al, 2001; Swail, 2001). Perhaps because of these perceived successes, the program was expanded so that, by 2010-11, there were 953 active UB sites across the country serving over 64,000 students for an annual

direct project cost of in excess of $314 million (U.S. Department of Education, 2010).[3]

With its extensive array of services, and at a yearly total cost of well over $5,000 per enrolled student, UB is 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; Harris, 2013; U.S. Department of Education, 2012).

Driven partly by its high profile and growing costs, the U.S. Department of Education funded and oversaw Mathematica Policy Research Inc. (MPR) to carry out a decade-long randomized trial. In its most recent report, the researchers concluded: "Upward Bound had *no detectable effect* on the rate of overall postsecondary enrollment or the type or selectivity of postsecondary institution attended" (Seftor et al. 2009, p.xv) and had "*no detectable effect* on the likelihood of earning a bachelor's degree or the likelihood of earning an associate degree" (Seftor et al. 2009, p.xvi). Regarding high school outcomes, an earlier MPR report drew similar conclusions: "Upward Bound had *limited or no effects* on total high school credits or grades . . . and *no effect* on honors and Advanced Placement credits, grades earned in high school or high school graduation" (Myers et al., 2004, p.xviii) (emphasis added in each quotation).[4]

These results from the experiment called into question UB's reputation as an exemplary program. Citing the first analysis of the randomized trial, the 2005 and 2006 White House budgets called for "zero funding" for UB (Field, 2007; Cahalan, 2009: Council on Opportunity in

[3] 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 percent (Harris and Goldrick-Rab, 2010).

[4] The two study summaries did report positive effects on postsecondary certificates and other effects for specific subgroups but, given the absence of effects on the average student for the majority of high school and college outcomes, these more positive findings drew little attention.

Education, 2012).[5] Haskins and Rouse (2013) recently quoted MPR's "no detectable effect" finding (p.3) and used this as their main evidence to justify a major program overhaul, including increasing competition among service providers, consolidating funding, and providing more flexibility in the use of funds. As additional support for their argument, they point out that a flurry of other more recent randomized trials have yielded promising findings for alternative college access programs such as providing information to students (Bettinger, Long, Oreopoulos, & Sanbonmatsu, 2009; Hoxby & Turner, 2013), college coaching (Bettinger & Baker, 2014) and facilitating the transition from high school to college (Castleman et al., 2012). Taken together, these studies suggest that some college access programs are effective, but that UB is not one of them.

The conclusions drawn from the experiment were controversial, however. Two key problems were identified with the sampling design and sampling weights that combined to significantly affect the conclusions: some observations carried up to 80 times as much weight as other observations, which greatly reduced precision. Moreover, the weights were probably incorrect because one of these heavily weighted sites likely belonged in a stratum that would have given it much less weight. Combined with the large negative effects on student outcomes in this same site, the estimates are both imprecise and biased. If the conclusions had been robust to various methods for handling this problem, the results might still have been convincing, but this

[5] Specifically, the MPR reports were cited by the White House Office of Management and Budget, which gave the program an "ineffective" rating. This rating was used to justify the elimination of funding not only for Upward Bound but Talent Search and GEAR UP as well (Cahalan, 2009).

turned out not to be the case, as MPR acknowledged at the end of its summary.[6]

Despite the imprecision, apparent bias, and lack of robustness in the average treatment effect estimates, there is still much we can learn from this experiment. We focus especially on effect heterogeneity for two reasons. First, from a methodological standpoint, we show that the problems that arose with the average treatment effects are less relevant in analyses of effect heterogeneity. Second, comparing the experimental sample with a nationally representative of the overall UB population, we find that the socioeconomic status of the UB experimental sample was much lower than the UB population. Therefore estimating population average treatment effects requires estimating separate effects for typical UB students.

The third motivation for effect heterogeneity analysis is rooted in the importance of Upward Bound as part of the *Great Society* and as a tool for reducing poverty and income inequality. We show that, in the population of UB students, family incomes have increased considerably over time. Moreover, in serving more advantaged students, the program has become less effective. While data limitations and non-representativeness of the experimental sample make it difficult to estimate effects by family income, students who are more disadvantaged on other dimensions clearly benefit at least as much as advantaged students. In particular, students in the experimental sample who are typically deemed ineligible, due to behavioral problems and low self-reported college expectations, saw a 10-percentage point increase in the probability of graduating high school and an equal-sized effect on receiving a college credential (especially certificates). These effect heterogeneity analyses have implications not only for efficiency of the eligibility requirements but for the generalizability of the estimates

[6] Later in the summary, the authors did write that "The lower robustness of the chosen sample design and the results from the extensive sensitivity analyses can be taken into account in determining the implications of the main findings" (Seftor et al. 2009, p.xviii), although they do not account for it in their own interpretation.

to the way UB is typically implemented.

Finally, we carry out cost-benefit and cost-effectiveness analyses of UB using various sets of impact estimates and evidence about the social and economic returns to various credentials as well as the direct costs of UB and opportunity costs of college. Due to the problems with identifying average treatment effects, it is unclear whether UB passes a cost-benefit test for typically eligible students. However, when well targeted, UB appears to be a cost-effective way to increase education outcomes compared with many program alternatives. Unlike estimates of the average treatment effects, the cost-effectiveness for typically ineligible students is robust to the ways we estimate treatment effects for subgroups.

We discuss the experimental design and data in Section II. Then, in Section III, we discuss the methodological issues, the ways we chose to handle them, and our main results for average treatment effects. Section IV examines effect heterogeneity, focusing especially on effects by student eligibility for the program and family background. The cost-benefit and cost-effectiveness analyses are in Section V.

II. Experimental Design and Econometric Methods

The UB experiment is formally known as the National Evaluation of Upward Bound and also, informally, as the Horizons study. We refer to it simply as the UB experiment. Below, we describe the UB experimental design, sampling weights, data and key variables, the econometric framework, and potential sources of bias.

II.A. Design and Sampling

Starting from a population of 568 operating UB projects, MPR identified a sampling frame of 395 UB sites managed by a local two or four-year college with at least three years of

hosting experience (Myers and Schirm, 1997).[7] The researchers then placed the 395 qualifying sites into 46 strata, selecting a stratified random sample of 67.[8] Within a given stratum, each site had an equal probability of being selected, but small and large sites as well as those administered by two-year colleges were oversampled to facilitate subgroup analyses (Myers and Schirm, 1997). The discussion of methods below shows that these sampling decisions have a significant influence on the analysis.

UB sites were required to have at least two applicants for each available opening. This oversubscription may have occurred naturally, or it may have been induced by pressure to comply with the experiment's over-subscription requirement, though MPR actively discouraged site administrators from relaxing admission standards. Congress has since banned the use of obligatory over-subscription in federally funded experiments.[9]

From within each site, MPR randomly selected individual students to be assigned to UB treatment. The final sample was composed of 1,524 treatment and 1,320 control students (Myers and Schirm, 1999). The mean site sample size was 42 students with a range of 4 to 96 (Myers and Schirm, 1997).

[7] Among other things the initial sample excluded sites that primarily served students with disabilities. Given our later findings about the positive effects of the program for students most at-risk of academic failure, this omission is noteworthy.

[8] The four variables used to create the 46 strata were: location of host institution (urban or rural), type of host institution (public four-year, private four-year, two-year), project size (small, medium, large), and historical racial composition of the project. MPR initially selected 70 sites; 11 sites declined and were replaced with 8 others yielding a total 67 (Myers and Schirm, 1997).

[9] Many legislators considered over-subscription 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).

II.B. Data and Descriptive Statistics

The USDOE's Office of Policy and Program Studies Service (PPSS) published the MPR analysis and provided the data for our analysis. Some of the data used by MPR were not included in the data files provided to us, as discussed below. To confirm that the files provided were otherwise the same as those MPR used, we checked the number of students and the descriptive statistics of key variables, which are essentially identical (see appendix Table A1). We also confirmed that our estimated average treatment effects (ATEs) are essentially identical to MPR's when we replicate their impact analyses.

All control and treatment students were required to complete background surveys and agreed to release their 9th grade transcripts prior to randomization. These surveys provide the main independent variables, summarized in Table 1. For simplicity, we use the same covariates as MPR as these choices seem to have a minimal influence on the findings.

Participants in the experiment were middle and high school students in grades 8-11 in the 67 experimental sites between May 1992 and March 1994. The applicants were mostly female and minority (67 and 72 percent, respectively) and living in households where parents did not attend college (94 percent "first generation") and had incomes below 150% of the poverty level (85 percent "low income"). As we will see later, UB students are also more socioeconomically disadvantaged than their non-UB counterparts. However, over time, UB students have become more like the typical U.S. high school student.

II.C. Econometric Framework

We identify UB effects from the randomization of students into control and treatment groups. Covariate adjustments are used to (potentially) increase statistical power.[10] We estimate the following model:

$$Y_i = \beta_0 + T_i\beta_1 + X_i\beta_2 + \varepsilon_i \quad (1)$$

where Y_i is educational outcome of student i, T_i indicates treatment assignment, and X_i is a vector of covariates (see Table 1). The key outcomes are high school graduation, high school GPA, college enrollment, and receipt of various types of college certificates and degrees. Outcomes are measured with data from five post-treatment survey waves, referred to as waves 1-5.[11]

In part to avoid switching models between GPA and dichotomous degree measures, we report Ordinary Least Squares (OLS) throughout. The results are substantively the same for the dichotomous outcomes when estimating the analogous version of (1) using probit. Robust standard errors are clustered at the site level in all models.

We study effects for various student subgroups: those with more disadvantaged family backgrounds and another group that has discipline problems and/or other characteristics that would typically make them ineligible. We therefore also estimate:

$$Y_i = \beta_0 + T_i\beta_1 + X_i\beta_2 + S_i\beta_3 + (S_i \cdot T_i)\beta_4 + \varepsilon_i \quad (2)$$

where the interaction term indicates whether the effects are different for the typically ineligible subgroup of students (S_i). Since these subgroup tests were not pre-specified, we follow

[10] Some researchers argue against using regression adjustment models when analyzing experimental data because the post-adjustment standard errors may be overly large or small. However this objection does not appear to apply to experiments such as this with sample sizes of over 1,000 (Freedman, 2008).

[11] The five survey waves were conducted in 1994–1995, 1996–1997, 1998–1999, 2001–2002, and 2003-2004 (Seftor et al., 2009). The last survey (wave 5) was completed seven years after the time that a ninth grader entering the sample pool in 1991 would have graduated on-time from high school.

Sanbonmatsu et al. (2006) and others, characterizing these as exploratory analyses.

II.D. Take-Up Rates and Treatment Contrast

As noted earlier, most of our estimates are intent-to-treat (ITT); however, we also estimate treatment-on-treated (TOT) effects and therefore briefly discuss the take-up rate. Twenty percent of students assigned to treatment never showed up for any UB program activity and we consider only this group to be untreated. Of those who attend at least some program activities, 37 percent typically exit UB after less than one year and about 65 percent of students leave the program prior to high school graduation (Myers and Schirm, 1999). Thirty-five percent of the treatment group completed all years of UB for which they were eligible (Myers and Schirm, 1999) with an average of 1.67 years of program participation (Seftor et al., 2009).

One criticism of the UB experiment is that 58 percent of the control group reported receiving access to other college access-related services, compared with 41 percent of the treatment group (Cahalan, 2009).[12] Participation in other programs is unsurprising, especially given the national growth in college access programs in the 1990s, precipitated in part by the perceived early successes of UB. Conversely, 20 percent of the treatment group received no treatment and 18 percent of the control group received college-related supplemental services that the treatment group did not (Myers & Shirm, 1997). This does not introduce bias, but it is important to be clear about the differences in services received by the control and treatment groups.

[12] About one in five of the students in these "other" programs (12 percent of the entire control group) were in other UB programs, especially "UB Math and Science," which operates as a separate program, but with similar services. It appears that these programs were available to the same students from other nearby sites, most of which were not part of the UB experiment.

II.E. Initial Identification Issues

In randomized trials, the primary threat to identification is usually attrition. The survey response rate in the second wave was 86 percent, but this declined with each round of survey and, by the fifth wave, the rate had dropped to 74 percent. We are primarily concerned with differential attrition between the control and treatment groups. Starting with the second wave survey, treatment group response rates were five percentage points higher than control group responses (88 versus 83 percent) and this difference persisted through the study termination date (76 versus 72 percent). In our preferred estimates, we combine waves, increasing the overall response rate from 81 and 74 for the high school and college outcomes, respectively, to 93 and 94 percent.[13] The control-treatment response differential is also reduced to just 3-4 percentage points by including these additional observations.

Comparing the control and treatment groups and baseline survey measures in Table 2, we find differences in only student educational expectations, which tend to favor the control group. The bottom of the table provides an *F*-test from a regression of treatment status on the full vector of covariates, which rejects the null. Nevertheless, the magnitudes of the differences seem relatively small.

[13] There is very little missing data in the baseline survey measures because filling out the survey was required in order to be in the experiment. Therefore, in our main analyses, we use only complete cases, but also considered multiple imputation (von Hippel, 2007) as a robustness check. This strategy involves imputing the missing data and then deleting the imputed observations that contained previously missing data on the dependent variable. Post-deletion, we analyzed the recombined data set using the same OLS estimation models as used under listwise deletion (Von Hippel, 2007). The outcome of the multiple imputation process is the imputation of missing covariate variables for cases where the dependent variable is known.

II.F. Sampling Weights and Non-Representativeness

Sites were selected via stratified randomization. With 67 sites selected from 46 strata, the vast majority of strata had only one site selected and some individual sites were used to represent large numbers of additional UB sites. According to Seftor et al. (2009), the sampling design was driven by decisions made by the U.S. Department of Education.[14] Also, while simple randomization was used to select students in most sites, some were allowed to stratify within sites so long as there were enough eligible applicants to support randomization within a stratum. The combination of stratification of site selection (46 strata) and stratification within some sites resulted in 339 total strata for random assignment, which were collapsed down to 192 due to empty strata (Myers and Schirm, 1999).[15] Within each stratum, students generally had an equal probability of selection.

Sampling weights are typically used in these cases so that the "analysis sample is representative of the target population" (Haider, Solon, and Wooldridge (HSW), 2013, p.2). As HSW (2013) note, however, such weighting is not always advisable for the estimation of causal effects. Weights can greatly reduce precision (Dickens, 1990) and may still yield inconsistent estimates of the population parameter, i.e., estimates that are as non-representative of the

[14] The authors write that "the Department of Education required the inclusion in the sample of substantial numbers of predominantly Asian, Native American, and Latino projects. Highly disproportionate and, therefore, unequal sampling rates were required to obtain such overrepresentation of these projects given that they are relatively rare in the universe of Upward Bound projects" (2009, p.A.8). Over-sampling particular groups in this way is a common reason for stratification and sampling weights, though the discussion that follows highlights the empirical trade-off, i.e., the loss in precision for the average treatment effects.

[15] The stratification procedures as well as decisions about the number of students per stratum were not clearly documented and this creates some problems later in trying to address issues with study design. Definitions for each of the 339 strata were not published. Also, 30 of the 339 strata ended up lacking either a control or treatment observation (Myers and Schirm, 1999). These strata were collapsed into the nearest neighboring strata.

population parameter as unweighted estimates (HSW, 2013).[16] The inconsistency can occur when there is effect heterogeneity that is not accounted for in the model.

HSW identify only one condition where the weighted estimates are consistent when there is unmodeled effect heterogeneity: when testing the differences in means (without covariates). Since the difference in means can only identify consistent estimates in a randomized control trial (RCT), this is also the only general situation in which applying sampling weights clearly identifies the population average causal effect.

This would seem to suggest that, setting aside other potential sources of bias such as attrition, estimating effects from the UB experiment with standard sampling weights is strictly preferred to the unweighted estimates. Unfortunately, the issue is less clear in this case because of other problems with the UB experiment sampling design. The first issue with the weighting scheme is the large variation in the probability of selection across sites. Students in five of the 67 projects were assigned approximately 43 percent of the weighted sample (Seftor et al., 2009). Across all students in the study, sampling weights for the 67 projects ranged from 1.9 to 184.8 (see histogram in the appendix Figure A1). The high end of this range is an order of magnitude larger than what is common in the literature.[17]

While the widely varying weights reduce efficiency, the HSW conclusions about consistency generally still apply. Again, however, the UB experiment is an exception because one site may have been placed in the wrong stratum. Evidence that this site number 69 was in the wrong stratum is based on additional inquiries made about that site after the experiment by staff

[16] More precisely, in the limit, the estimates do not converge to the *population* average treatment effect.
[17] While there seems to be no standard rule of thumb, our search of the literature yielded maximum weights usually in the 6-11 range.

of the U.S. Department of Education (Cahalan, 2009).[18] Unfortunately, these inquiries were apparently not made in other sites.

To the degree that the wrong stratum problem introduces any non-representativeness, the very large weight given to site 69 compounds the problem. Employing the sampling weights as given not only reduces precision considerably, but also makes it impossible to identify the ATE for the target population. To see this more formally, we take a simple case in which the population ATE, β^*, is a weighted average of effects from two types of sites (i.e., sampling strata), A and B, such that $w_.^*$ is the true weight:

$$\beta^* = w_A^*\beta_A^* + w_B^*\beta_B^* \tag{3}$$

where the true effects are constant within site types. However, the weights are measured with error so that $\widehat{w}_. = w_.^* + e_.$. Substituting the measured weights for the true weights and re-arranging yields:

$$\beta^* = (\widehat{w}_A\beta_A^* + \widehat{w}_B\beta_B^*) - (e_A\beta_A^* + e_B\beta_B^*) \tag{4}$$

If the errors in the weights were independent of the true site effects, then the second term would equal zero and the expectation of (4) would be β^*. To see why, note that $E[e_.\beta_.^*] = \mu_e\mu_{\beta^*} + Cov[e_., \beta_.^*]$. If $\mu_e = 0$ within each type of site, then $\mu_e\mu_{\beta^*} = 0$, but the covariance term remains. In general, it might be reasonable to assume that $Cov[e_., \beta_.^*] = 0$.[19] However, the evidence

[18] Specifically, it is argued that site number 69 does not offer on-campus housing, had just begun to offer four-year degrees, and students applying for UB with this site were only interested in vocationally oriented training that might lead to certificates and two-year degrees (Cahalan, 2009). That interpretation is consistent with the results shown later that reducing the weight on site 69 increases the effect on BA completion.

[19] It is also worth noting that if one site is in the wrong stratum, then there are errors in all the other weights. This is because the weights must satisfy the condition $\sum_N w_.^* = N^*$ where the weights are summed over *N*, the number of observations in the sample and $N^*$is the true population size the sample is meant to represent. By substitution, in a two-stratum case, $\sum_N w_.^* = (\widehat{w}_A + e_A) + (\widehat{w}_B + e_B) = N^*$. Since N^*, $\widehat{w}_A$, and $\widehat{w}_B$ are all known, $e_A$and e_B have to cancel for this condition to hold. Again, however, any correlation between the weight errors and true effects remains speculative.

presented above suggests this is not the case in the UB experiment: the weight attached to site 69 has a (probably large) positive error and a large negative treatment effect (see Figure 1). It is possible that the errors in the weights for the other sites are unrelated to the site treatment effects, or even that the correlation in the other sites is in the other direction, offsetting the bias due to site 69, but this is purely speculative. In short, the combination of widely varying weights and the potential wrong stratum problem with site 69 means that a strong assumption is necessary to identify unbiased estimates for the target population in this experiment.

We try several approaches to deal with this problem. The most obvious potential solution is to place site 69 in the correct stratum and re-calculate the weights, but this is difficult to carry out because there are no other sites in the site 69 stratum to represent that part of the population, compounding the problem with the sampling design.[20]

An alternative is to "trim" the weights, i.e., reduce the extreme weights to some fixed level to increase precision (Potter, 1988; Potter, 1990; Liu, Ferraro, Wilson, and Brick, 2004; Pedlow, Wang, Yongyi, Scheib, and Shin, 2005; Chowdhury, Khare, and Wolter, 2007; Elliot, 2008). At the extreme, trimming all weights to a maximum of 1.0 (or any other single number) is the equivalent of ignoring the weights altogether.

One version of trimming involves estimating effects at different trimming levels and identifying the trimming level with the smallest mean-squared error (MSE) of the treatment effect (Cox & McGrath, 1981; Potter, 1988). In this case, the MSE-minimizing method yields trimming at the 0-5th percentile of the weight distribution for high school outcomes and post-

[20] Even if there were additional sites in the site 69 stratum, we do not have enough data to correctly move sites to different strata.

secondary enrollment (i.e., similar weights for all observations) and at the 95-100th percentile (i.e., minimal trimming) for other college outcomes.[21]

With multiple outcomes, the disadvantage of MSE-minimized trimming is that the weighting scheme has to be re-adjusted for every estimate. A simpler, though more ad hoc, approach involves trimming all the weights at the same percentile. Some simulation evidence suggests that trimming up to the 30th percentile and down to the 75th percentile reduces the mean squared error of the estimate in other settings (Asparouhov & Muthen, 2005).[22] Taking the 75th percentile of the weights in this case reduces the maximum sampling weight from 184.8 to 15.3. Another solution is to simply drop the problematic site (Cahalan, 2009), which implies trimming the weight to zero, but only for this single site.

All of these approaches reduce the weight attached to site 69, but, given the earlier discussion of errors in the weights, it remains unclear whether this increases or reduces bias (in relation to the population parameter). The choice of sampling weights and trimming might be irrelevant if the results were similar across the various options, but this is not the case as our general conclusions change when shifting from the unadjusted weights to almost any form of weight trimming.

In most studies of policy and program effects, the sample is chosen for convenience. A city or state implements a program and researchers attempt to identify causal effects that are representative of some (usually unstated) set of similar sites. The discussion here highlights how

[21] This varies somewhat among the various post-secondary outcomes, but all are above the 95th percentile. The fact that the optimal trimming level varies by outcome measure is due in part to the fact that the correlation between the site weight and site effect varies by outcome.

[22] One justification for this approach is highlighted by the earlier text indicating that the weights have to sum to the population size. If negative errors in some sites are offset by positive errors in other sites, the variance of the measured weights will be larger than the true weight distribution. Therefore trimming at the top and bottom may move the distribution of weights closer to the true distribution.

identifying effects that are both unbiased and representative of a specific population, as in the UB experiment, is more challenging and requires additional assumptions, even in a randomized trial. In the next section, we identify an additional factor, other than the sampling weights, that calls into question the representativeness of the ATEs estimated using the unadjusted sampling weights.

III. Average UB Effect Estimates

III.A. Preferred Model

Table 3 shows our estimated ATEs with the available data on six outcomes: high school graduation and GPA, college enrollment (i.e., any type of certification or degree-seeking enrollment), and completion of any college credential, two-year/associate degrees and above (AA+), and four-year/bachelor degrees and above (BA+). For each outcome, we report estimates from six specifications, varying the use of covariates, weights, and inclusion of site-specific intercepts.[23] All tables report robust standard errors clustered at the site level for all estimates.[24]

Column (6) suggests that UB increased high school graduation, college enrollment, and certificate and BA completion each by about 3-5 percentage points. Absent UB, approximately 38 out of 100 high school students completed some post-secondary education. For students enrolled in UB, that number rose to 43 students, a relative increase of 12 percent.[25] These are our preferred estimates because they use the sampling weights that account for both bias and

[23] There are two main potential rationales for including site-specific intercepts: (a) to increase precision; and (b) to account for site variation in the relationship between missing covariate data and education outcomes.
[24] This is implemented with the svy command in Stata.
[25] It is possible that some of the estimates are significant because of the number of comparisons. Bonferroni adjustment involves multiplying the p-values by the number of different outcomes (in this case six). Therefore, as a general rule, here and throughout the paper, the estimates with a single asterisk ($p<.05$) are insignificant after the adjustment, but those with two or more asterisks are still significant after the adjustment.

efficiency considerations.

But these estimates are sensitive to the estimation procedure. In particular, comparing across the columns, the estimates are almost always more positive when the weights are trimmed. The extreme case is when no weights are applied (columns (1) and (2)), in which case site 69 is three percent of the sample. The trimmed weights give site 69 a much greater contribution and the unadjusted sampling weights give the site the largest weight (26 percent of the weighted sample). Clearly, the attention given to site 69 is well deserved. While there is at least one statistically significant effect in each specification, there is also at least one insignificant estimate in every row (i.e., for every outcome measure).[26]

It is this lack of robustness that has generated so much controversy. As in prior studies of school vouchers (Krueger & Zhu, 2003; Mayer et al., 2002) and school competition (Hoxby, 2000; Rothstein, 2007), seemingly slight changes in methods yield substantively different conclusions. This naturally reduces confidence in the conclusions, especially in this case where none of the estimates is clearly preferred to the others.[27]

III.B. Direct Comparison with MPR

To highlight the role of the sampling weights in the MPR analysis more directly, the first two columns of Table 4 are copied from the most recent MPR report. Column (A) is their preferred specification. Column (B) shows the specification that we can most closely match with our data, while Column (1) shows our replication of that specification. Comparing (B) and (1),

[26] Comparing columns (3) and (4), it is clear that adding site-specific intercepts makes almost no difference therefore we do not report specifications with site effects going forward.

[27] Recall that $E[e_{\cdot}\beta_{\cdot}^{*}] = \mu_e\mu_\beta + Cov[e_{\cdot},\beta_{\cdot}^{*}]$. Changing the weights changes the errors and thus the covariance between the error and the subgroup treatment effect.

the point estimates and significance levels are very similar for all outcomes.

We cannot replicate MPR's preferred specification (column A) because MPR had access to additional measures of college outcomes from the National Student Clearinghouse (NSC) and Student Financial Aid (SFA) files. For college outcomes, MPR's preferred approach was to use a combination of the fifth survey wave, NSC, and SFA data, whereas we are forced to use surveys only.[28] While the NSC is now considered the best source of data for post-secondary enrollment and completion, the coverage has always been lower in the two-year sector and was apparently even lower in 2006 when the NSC data were collected. MPR reported that six of the 67 UB sites were not participating in the NSC at the time the data were pulled and another 10 sites were not participating at the time that students might have first graduated from college.[29] The SFA data are likely to be invalid because many students who attend two-year non-profit colleges never apply for aid.[30] While the NSC and SFA both have limits, there is no reason to think measurement error would be systematically different between the control and treatment groups (Goldrick-Rab & Harris, 2010).

To test the potential influence of omitting the NSC and SFA data, we turn to additional estimates reported by MPR in their extensive appendices (not shown in Table 4).[31] Specifically,

[28] Briefly, MPR's preferred method was to designate a student at "enrolled" if there was any evidence of enrollment in either the fifth wave survey, NSC, or SFA; a student "completed" if there was any evidence of completion in either the fifth wave survey or the NSC; all others are non-enrolled or non-completers, respectively. We took the same approach in combining survey waves. This coding procedure precludes missing college outcome data in the MPR analysis, while leaving a 6-7 percent missing data rate in our analysis.

[29] Students were randomly selected for UB in 1992-93 and 1993-94 when the typical student was a freshman in high school. This means on-time high school graduation would have been in 1996 and 1997. For students going directly on to college full-time, they might reasonably have received a two-year degree in 1998/1999 or a four-year degree in 2001/2002. Thus, it does not appear that NSC would have captured enrollment in institutions that began NSC participation after 2002.

[30] For this reason MPR only used the SFA data to change students from credential non-completers to completers.

[31] At first, comparing the two MPR estimates in columns (A) and (B), it appears that the NSC/SFA omission inflates the estimates. But this is not the right comparison because (B) uses only a single survey wave and our preferred estimates combine all survey waves.

we compared two MPR estimates that both use the 5th survey, NSC, and SFA. The point estimates are smaller in magnitude, and smaller than MPR's preferred estimates, when the data also include the third and fourth survey waves. Since we are also using all the survey waves, this provides additional evidence that our estimates are conservative and biased toward zero.[32]

The last three columns in Table 4 are provided to show, step by step, how the results change as we proceed from MPR's preferred specifications to our own. This reinforces the lack of robustness to sampling weight changes found in Table 3. The conclusions are more positive than MPR's when the sampling weights are handled in almost any other way. It is for this reason we conclude not that there was "no detectable effect" but that no clear conclusion can be drawn about the ATEs. Fortunately, this same problem does not arise in the analysis of effect heterogeneity.

IV. Effect Heterogeneity by Eligibility

Two additional concerns have been raised about UB and the UB experiment that call for effect heterogeneity analyses. First, the (weighted) experimental sample may have differed from UB population (Cahalan, 2009), creating yet another threat to identification of the population ATE under business-as-usual. Second, there have been suggestions since UB's inception that local eligibility requirements might have targeted the program to students who benefit least from treatment exposure (Greenleigh, 1970; Fields, 2007). Therefore, we start by using additional data to compare the experimental sample to the UB population and study effect heterogeneity by eligibility status.

[32] It is important to note in these comparisons that, when all the survey waves are used, only 6-11 percent of observations have missing college outcome data. Therefore, the addition of the NSC and SFA data alter the measured outcomes minimally.

We carry out two types of analysis to understand whether the composition of the UB population has changed over time in ways that may have reduced benefits overall and particularly for less advantaged students. Using data from outside the experiment, we test for trends in the socio-economic status of the UB population over time using multiple national UB samples. Then, we carry out additional effect heterogeneity analyses with the experiment data to determine whether the trends in socio-economic status have changed program efficiency and/or the degree to which UB helps reduce college achievement gaps.

Finally, we consider whether effect heterogeneity by student subgroup might reflect effect heterogeneity across UB sites. This also allows us to test whether specific elements of UB programs drive effectiveness.

IV.A. Eligibility of the Experimental Sample versus the UB Population

Critics of the UB experiment argue that the sample was not representative of the UB population, perhaps because of the need for lottery over-subscription (Cahalan, 2009). To test this, Tables 5A and 5B compare the experimental sample of UB participants with two other national surveys, the National Educational Longitudinal Study of 1988 (NELS:88) and the High School Longitudinal Study of 2009 (HSLS:09), both of which identify UB participants. These two additional data sources are from somewhat different periods in students' high school careers, and in some cases use somewhat different measures, so we report different types of comparisons in the tables and focus on only the most comparable measures.[33]

Chronologically, the experimental sample falls in between the NELS:88 and HSLS:09, so we would expect the UB experiment data to fall in between the two, closer to the NELS:88. This

[33] See the Appendix for more detail about variable definitions and comparability. Also, while Table 2 is focused primarily on capturing students' absolute level of socioeconomic advantage, comparisons between UB students and the overall student population are provided in the Appendix.

is not what we see, however. Compared with both nationally representative UB samples, the experimental sample has a much larger percentage of students below the poverty line and much lower levels of parent education. Though MPR did apparently go to great lengths to keep eligibility the same, it does not appear that the sites complied as the experiment was implemented.

At the same time, the sites apparently maintained their academic standards. The experimental sample had a somewhat lower GPA than the NELS:88 UB students, but almost identical percentages of students had taken algebra (neither is statistically different between the samples). This finding suggests that sites may have attracted more students to apply for UB by finding additional students that met the usual UB profile on academics, but who might not otherwise have applied to the program. For example, UB might not have been widely advertised to students by UB site directors before the experiment. Students with higher socio-economic status might have been more actively looking for such programs and found UB on their own. With more active recruitment, the experimental sample might have reached a broader audience, including lower socio-economic status students who were not actively searching for college access programs. This theory is speculative, but does fit the pattern in these data and prior research.[34] Whatever the reason, the differences between the MPR sample and the other national samples are large enough to conclude that something was clearly different in the way eligibility was operationalized in the experiment compared to business-as-usual.[35]

Additional data from the UB experiment allow us to separately identify specific students

[34] There is evidence from prior studies of education and other social programs that individuals most in need are less likely to seek out the programs that might serve them.
[35] There was a change in official program eligibility rules in 1995, but this was after the experiment and the direction of the rule changes is inconsistent with the patterns in Table 5.

who may have ended up in the experiment even though they would have been typically excluded.[36] Specifically, since sites have discretion over some eligibility criteria, we used data from grantee surveys that described the eligibility decisions made by site administrators. The grantee survey that MPR administered in 1992 (prior to the experiment) showed that a two-step process was used to construct the list of eligible candidates once federal eligibility had been established (Moore, et al., 1997). As a first step, 95 percent of project directors relied on positive recommendations from feeder school educators and staff before considering a student eligible to receive local UB services. In the second stage, 86 percent of project directors then used one or more screening criteria to identify ineligible students.

The top seven reasons used by site directors to render a student ineligible were: (1) no specific interest in college (46 percent); (2) history of emotional or behavioral problems (41 percent); (3) history of drug or alcohol abuse (41 percent); (4) gang activity (34 percent); (5) record of disciplinary actions (34 percent); (6) GPA above a specified minimum (32 percent); and (7) GPA below a specified minimum (26 percent). Sixty-two percent of site directors immediately excluded any student who reported any history of behavioral problems (Moore et al., 1997).

As the basis of comparison, we also considered the baseline data from the student survey in which students reported measures of their attendance, tardiness, suspensions, expulsions, breaking school rules, arrests, forced transfers to other schools or juvenile facilities, and

[36] A few other explanations are possible. First, the MPR sampling frame excluded certain types of UB sites that serve different types of students. As noted earlier, there were 568 sites, of which 395 ended up in the sampling frame (i.e., 173 sites were omitted prior to sampling). It is not clear whether the students served are systematically different in these sites; even if the omitted sites do differ, this would still mean the experimental sample is non-representative. Also, the NELS and HSLS might be non-representative of the UB population in their respective time periods, but there is no reason to believe this is the case.

expectations of high school and college graduation.[37] We set thresholds for each of these variables based on the grantee survey (see Appendix Table A2) and then identified students who failed on any one of the criteria. The most common single reason for a student to be ineligible is low college expectations.

In some cases, the grantee and student survey questions line up well (e.g., students were asked whether they had been in juvenile detention and this partly aligns with "history of disciplinary actions" in the grantee survey). In other cases, there are not even partial matches (e.g., the student survey never mentions the drug and alcohol use, gang activity, or prior GPA, though the grantee survey suggests that site administrators consider these factors). Also, students might under-report negative behaviors such as in-school suspensions.

We identify as ineligible only those students who clearly reported any behavior inconsistent with grantee-reported eligibility criteria. We also carried out the above general approach in two different ways: (a) using the *average* grantee survey results, we established a single set of blanket criteria and applied that to students in all sites; and (b) using the *site-specific* grantee survey results, we applied separate criteria to each site. Approach (a) is preferable to the degree that measurement error in the grantee surveys is large relative to the cross-site variance in actual criteria, while (b) is preferred when measurement error is relatively small. The fact that we only identify students at the extremes of the eligibility criteria, combined with the above measurement problems, means that we are likely under-stating ineligible students, though it is unclear by how much since the included eligibility measures are likely correlated with the

[37] The questions in the student survey included the following: "During the last school year how many times did each of the following things happen to you?" Responses included "I got in trouble for not following school rules," "I was put on in-school suspension," "I was suspended or put on probation from school," "I was transferred to another school for disciplinary reasons," "I was arrested," and "I spent time in a juvenile home/detention center."

omitted ones.

As additional evidence of our eligibility measure, we tested whether the ineligibility rate from the above decision rules lined up with the overall rate of ineligibility reported by grantees under business-as-usual. In conjunction with the grantee survey report, MPR also collected survey data on the experiences from a representative sample of UB feeder schools showing that, 43.4 percent of applicants were screened out annually during the three years that preceded the experiment (Moore, et al., 1997). Since students should be screened out prior to randomization under business-as usual, we would expect to see very few students in the sample with traits that would typically make them ineligible. This is not what we find. Though we were conservative in identifying students as ineligible, 10-23 percent of students appear ineligible using the various criteria.

The number of apparently ineligible students, combined with the differences between the experimental sample and other nationally representative samples, suggest the UB experimental sample and effects do not represent typical program operations. We therefore go further and estimate effects by eligibility status.

IV.B. Effect Heterogeneity by Eligibility

To test whether the change in UB population and discrepancies from business-as-usual influenced program effectiveness, we re-estimated the Table 3 ATE specifications for the various eligibility subgroups. The Appendix discusses more formally the possible scenarios for bias in the subgroup effects building on equations (3) and (4), but the bottom line is that we cannot determine whether any particular is unbiased given the problems with the sampling design.

We do still have to be concerned about robustness, however. The various trimming alternatives change the weight errors and therefore may change the mean error (μ_e) within strata

and will change the covariances between the errors and the site treatment effects ($Cov(e_{\cdot}, \beta_{\cdot}^{*})$). If the results were robust, then this would suggest the errors are small.[38] If they are not robust, then we run into the same problem as with the ATEs.

Tables 6A and 6B show separate effects on typical UB students versus those who are typically ineligible. For typical UB students, the effects are similar to our preferred specifications from Table 3. Again, the effects on high school graduation and BA degree completion are all statistically significant in the majority of specifications, and the enrollment effects are sometimes significant. Since these estimates are for typical UB students, these are our best estimates of the ATEs under business-as-usual.

The story is noticeably different in Table 6B for typically ineligible students. Given that ineligible students tend to have more academic and behavioral problems, these students should be closer to the margin of receiving lower-level academic credentials. For example, it might be difficult to get a student who has been assigned to juvenile detention to graduate with a BA degree, but high school graduation and certificates are more feasible and also have substantial economic returns. This is what we find. There are no BA effects, but possibly large effects of about 10 percentage points on high school graduation and college credentials. The effect on certificates (and above) is always in the range of 6-12 percentage points across all specifications and often statistically significant. The high school graduation point estimate is similarly consistent, falling in the 6-11 point range, but these are significant in only two of the six specifications. The sample sizes in this table are generally quite small, ranging from 261 to 630, which keeps the standard errors large.

[38] Changing the weights changes the errors and covariances with the effect estimates, so if the point estimates do not change, the errors are probably small. See the Appendix for more details.

Of the 36 estimates in Table 6B, 32 are more positive for the typically ineligible students, far more than would be expected by chance.[39] The *p*-values at the bottom of Table 6B are based on the interaction term shown in equation (2) and these show that five of the 36 coefficient pairs are statistically different from one another at $p<0.05$ and another three are significant at $p<0.10$. In all of the eight cases where there are at least marginally statistically significant differences, the effects are more positive for the typically ineligible students. So, while many of the individual coefficients are not quite statistically distinguishable, there is a clear pattern.

The results are robust when we use the site-by-site criteria as the baseline method (available upon request). However, the differences in point estimates do appear to be sensitive to the choice of sampling weights. In some cases the shift to unadjusted weights increases the differences and in other cases it reduces them. Nevertheless, the general conclusions about costs and benefits for the various subgroups do turn out to be robust, as we show in section V.

IV.C. Trends in UB Participant Family Background and Academic Risk

The above analyses address differences between the UB experimental sample and business-as-usual. In this section, we consider whether the business-as-usual UB population has changed over the past several decades and how such changes may have influenced program effectiveness.

There are good reasons to expect a priori that the UB population has shifted over time. First, program eligibility criteria have changed to allow site operators more flexibility. In its early history, the federal government required that academic risk be one of the requirements for

[39] Though college expectations are part of the eligibility criteria, two of the three estimates omit college expectations because this measure is easily manipulated. For instance if a student is told that program acceptance is contingent upon an expectation of future college enrollment, they may misreport their true expectations. The other measures are based on past behavioral, which is generally verifiable by the site administrators.

UB eligibility. The 1976 RTI noted that 75 percent of participants were required to be low-income students with high college potential who needed academic and/or psychological support. Starting in 1995, however, there was no longer any requirement to admit students based on academic risk and, apparently as a result, the number of students classified as at-risk dropped by half, from 51 percent to 25 percent (Moore et al., 1997).

The service offerings have also changed. The 1976 RTI report noted that the interventions were focused primarily on remedial instruction, but by 1993, only three percent of the UB projects focused exclusively on remediation and 73 percent lacked any remedial component (Moore et al., 1997; Myers and Schirm, 1999). By 2000, there was also a shift towards offering a greater number of math and science courses, more advanced math, sciences, and foreign language courses (Moore et al., 1997; Cahalan, 2004).

To test whether these changes in policy led to changes in the UB population, we turn again to the NELS:88 and HSLS:09. The two data sources are from somewhat different periods in students' high school careers, and in some cases use somewhat different measures, so we report only the most comparable measures.[40]

As expected, in both 1998 and 2009, non-UB students were more advantaged than UB students. However, over these two decades, the percentage of UB students whose parents had completed an AA degree or BA degree (or higher) more than doubled. Similarly, the percentage below 130% of the poverty line dropped by five percentage points (both are significant). These trends are not due to changes in the overall population. Table 5 also shows simple difference-in-difference (DD) $((\text{Non-UB}_{09} - \text{UB}_{09}) - (\text{Non-UB}_{88} - \text{UB}_{88}))$ such that, for a trait where higher values indicate more (less) advantage, a negative (positive) DD estimate means that the UB

[40] See the Appendix for more detail about variable definitions and comparability.

population became relatively advantaged over time.

The DD point estimates of all ten measures of parent education and income suggest that UB students have become more advantaged over time. Eight of the ten are at least marginally significant and are arguably large in magnitude. For example, there was a 20-percentage point absolute gap between UB and non-UB student in whether parents had at least an AA degree, but this gap was cut to 13 percentage points by 2009. Put differently, the education levels of parents of UB students in 2009 were about as high as those of *non-UB* students in 1988.

The relative family income advantage of non-UB students also dropped by almost $7,000 between the two time periods.[41] Overall, it appears the policy changes discussed above helped induce a large upward shift in the socioeconomic status of UB participants, making them look much more like the average student in the population.

IV.D. Effect Heterogeneity by Family Background and Predicted Outcomes

It is unfortunately difficult to test whether the program became less effective because it was serving more advantaged students because almost all of the students in the sample are low-income (recall that the experimental sample is more disadvantaged than the UB population). Instead, we estimated effects using equation (2) where the subgroups are defined by the combination of low-income and first generation in college status. The vast majority of students are in both categories, enough students are in only one of those categories to identify somewhat precise estimates.

[41] Based on academic measures, the changes appear smaller and are not statistically significant. One potential reason is that these measures come from post-treatment data collection. If the program offerings had remained constant then it might have been reasonable to assume that the program impacts cancel out (we are only comparing UB students to other UB students across the samples). Since the program offerings did change and there is effect heterogeneity, it is not clear whether this is a reasonable assumption. This is another reason to focus on student socioeconomic status. It is somewhat implausible that UB could have affected these measures.

In Table 8, the reference group includes students who are both low-income and first generation. The coefficients on low-income-only and first-generation-only therefore indicate whether the effect gets larger or smaller when students have only one of these characteristics. In some ways, the results are consistent with what we found for eligibility. For AA and BA completion (as well as GPA), students who are low-income-only seem to benefit more than students who are also first generation. All but three of the 25 point estimates are positive and seven are statistically significant. They are also large in magnitude, e.g., increasing BA degree completion by more than 10 percentage points in every model. Also, there are signs that students who are first generation-only (not low-income) benefit more in terms of credentials, though these interactions are not significant.

This is not the case for lower-level education outcomes, however. For high school graduation and college enrollment, the estimates are still positive and significant for students who are both low-income and first generation. These patterns reinforce the theory that less disadvantaged students are more commonly on the margin of receiving college degrees while more disadvantaged students are on the margin of receiving high school diplomas and enrolling in college. College access programs will be more effective along the margins where students are at baseline and these margins are correlated with student socio-economic status.

The above evidence suggests that effect heterogeneity may have more to do with the specific credential margins than with socio-economic and behavioral advantage per se. To more directly test this theory, we considered the growing literature that estimates effects based on the predicted outcomes of participants (Hansen, 2008; Sanbonmatsu, Kling, Duncan & Brooks-Gunn, 2006). As Abadie, Chingos, and West (ACW, 2013) show, estimating effects by predicted outcomes is difficult to apply in randomized trials because we generally only have access to the

control and treatment groups when making the predictions. Including the treatment group is problematic because the predictions are conflated with the treatment effects, but limiting to just the control group (and applying the model as an out-of-sample prediction to the treatment group) creates a different problem because the prediction model has better statistical fit for the control group relative to the treatment group. This disparity in fit inflates the differences in effect estimates between the lower and higher predicted outcome groups. We adapted the ACW proposed methods to work with the UB experiment sampling design. Unfortunately, the first-stage predictions held little explanatory power and the patterns were sensitive to the number of bins students were placed into (e.g., terciles versus quartiles). Also, because there is uncertainty in both the prediction and the treatment effect (accounted for with bootstrapped standard errors), the estimates are very imprecise.

Again, since there are very few non-low-income students in the sample (and no continuous measure of income in the data), these analyses do not allow us to test whether the increasing level of socio-economic advantage in UB has increased program efficiency, nor can we provide convincing evidence about effects based on predicted outcomes. But the results do reinforce one earlier pattern: the benefits for advantaged students (with high predicted outcomes) are concentrated in BA completion and the benefits for less advantaged students are in high school graduation and lower-level college credentials.

IV.E. Site-Level Effect Heterogeneity

Another reason we might find different effects for more disadvantaged is that certain groups of students may be, coincidentally, served by UB sites that are more effective. In that case, the actual causal mechanism may relate to effect heterogeneity by site rather than by

student group. Analysis of site effects is also helpful for understanding what specific elements of programs may be most important to program effectiveness.

Figure 1 plots site effects and confidence intervals (note the large negative effect on site 69). This type of variation is not unusual or surprising, especially in light of the small samples in most of the sites. More important is whether this variation can be explained and therefore provide some insight into what makes some sites more effective than others. We turn again to the grantee survey and use measures that describe specific program characteristics (e.g., number of specific services and staff-participant ratio), site characteristics (e.g., two-year versus four-year sector), and contextual factors (e.g., location in an urban area).

To isolate the role of site effects from student ineligibility, these regressions omit ineligible students. There appears to be no relationship between ineligibility and site effects in any of the models, suggesting that the larger effects for ineligible students identified in Tables 6A and 6B are not driven by site-level effect heterogeneity.[42]

Consistent with the notion that UB benefits more disadvantaged students, we find that treatment effects are much smaller for suburban sites compared to omitted rural category (urban sites are no different). One potential reason, though speculative, is that low-income families living in or near suburban colleges experience better school resources and more positive peer and neighborhood spillovers that diminish the value of formal college access programs.[43]

We counted the number of specific services provided by UB sites and divided them into five groups using the grantee surveys categorization: college preparation (e.g., test preparation and college campus visits), work/career preparation (e.g., visits with local employers), self-

[42] We implemented this with the "mixed" command in Stata, which treats the site effects as random effects. See table notes for more detail.
[43] It is also worth noting that the sample includes only three suburban sites (and 80 students).

awareness seminars (e.g., health and nutrition), field trips (other than campus visits), and counseling (mostly non-academic). These are generally unrelated to site effects, although there is some indication that college preparation helps get students into college, but that this effect fades out with college credentials. Also, work preparation services may actually reduce education credentials, perhaps because students become more interested in working than continuing their formal education.

A key finding of the preceding analysis is that students with behavioral problems seem to benefit more from UB than other students. Reinforcing this conclusion, we find that sites that have performance requirements for continued program eligibility after initial UB entry experience smaller program effects. (Unfortunately, there is no additional information about the nature of these performance requirements.)

Staff-participant ratios, program size, and staff experience do not predict program effectiveness. This is consistent with the general findings about teacher effectiveness, which seems to be associated less with measureable educator characteristics than unobservable factors (Hanushek et al., 2005; Harris & Sass, 2011).

V. Cost-Benefit Analysis

The debate about Upward Bound, and the analysis so far, has focused on whether or not there is a detectable effect. But now that we see evidence of statistically significant effects, the question becomes, are the benefits worth the costs? Also, in the subgroup analysis, how do the benefits that accrue to ineligible and disadvantaged students through the acquisition of lower-level credentials compare to the benefits for typically eligible/advantaged students who benefit in higher-level credentials? To answer these questions, it is necessary to estimate and combine the welfare effects for each credential.

We use the various point estimates for the effects of UB on credentials from Tables 3 and 4, adjusting them using instrumental variables to estimate treatment-on-treated (TOT), or the Local Average Treatment Effect (Imbens & Angrist, 1994). These are more relevant for a cost-benefit analysis than the previously reported ITT effects.[44] We also took two different approaches to the point estimates, first taking all the UB point estimates at face value, ignoring statistical significance and, second, setting the insignificant estimates to zero.

While there is a vast literature on the return to a year of education and many studies that examine returns to high school graduation and bachelor's degrees, fewer studies examine the return to two-year degrees and credentials. We report a range of estimates based on Jepsen et al. (2014), Kane and Rouse (1995), Marcotte, Bailey, Borkoski, and Kienzl (2005), Levin, Belfield, Muennig, and Rouse (2006) and Tyler, Murnane, and Willett (2000), taking their preferred estimates and adjusting them as necessary so that high school dropouts were the reference group in each case. Averaging across these studies, our baseline estimates of the return to various credentials are: BA (2.25 times high school dropout earnings), AA (1.82 times), certificate (1.58 times), and high school degree (1.40 times).[45] The estimates overstate the social utility of education due to sheepskin effects (Weiss, 1995), but this is at least partially offset by under-stating external benefits (Wolfe & Haveman, 2002) and the fact that credentials improve matching of employees to employers and therefore increase the efficiency of the labor market.

Additional key parameters for this analysis include the economic return to credentials,

[44] The TOT estimate is the effect if there were zero "no shows." TOT= ITT/(1-ns) where "ns" is the no-show rate, so the TOT is always at least as large as the ITT. The no-show rates for the overall sample and ineligibles-only are 20 and 25 percent, respectively. The TOT is more appropriate for a cost-benefit analysis because this aligns with costs per participant served. That is, no resources are used to serve students who never receive services.

[45] We also assign zero return to higher education that does not result in a credential. Though some research suggests that there are such returns (e.g., Kane & Rouse, 1995), we cannot identify years of education in these data and instead make the more conservative assumption.

discount rate, and productivity growth. In our baseline estimates, costs and benefits are both discounted at a rate of 3.5 percent (Lipscomb, Weinstein, & Torrance, 1996; Moore et al., 2004; and Muennig, 2002). Following Levin et al. (2006), future productivity growth is assumed to be approximately 1.5 percent per year. The returns are estimated separately for each year in the workforce post-schooling (up to age 65).

On the cost side, we account for the direct costs of UB at about $5,000 per year, the direct cost of college for those who attend, and the opportunity cost to students of attending college.[46] If UB keeps students in college longer, this creates direct college costs, which we estimate based on Harris (2012) and Johnson (2009). We assume annualized costs of $9,700 per year for a certificate or AA degree program versus $11,280 for a BA degree program.[47]

To make the calculations concrete, consider a student who shifts upward from an AA to a BA as a result of UB. In the cost-benefit analysis, this is would increase benefits in proportion to the difference in the return to AAs and BAs (see above baseline values) and increased costs due to the larger annual costs of BA degree programs and the additional opportunity cost of spending more years in college.

Because of the apparent inclusion of ineligible students in the experiment, our best estimates of the net benefits for UB under business-as-usual come from the third and fourth columns of Table 10. These show consistently large and positive net benefits under a range of assumptions. In contrast, the MPR estimates hover around zero with a mix of positive and

[46] This somewhat under-states costs because while it counts the UB costs during high school, it does not count the marginal cost of the regular programs students experience in high school when they do not drop out.

[47] The costs of two-year and four-year colleges may seem too similar, but note that public subsidies represent twice the share of revenue in the two-year sector versus the four-year sector (U.S. Department of Education, 2012), so that tuition differences over-state cost differences. Also, the reported cost of four-year college often includes room and board, which does not apply to two-year colleges. We exclude room and board because most of these costs need to be incurred by adults regardless of whether they are in college. In 2012-13, the average room and board charges totaled $5,296 (constant dollars) at four-year colleges (U.S. Department of Education, 2014).

negative estimates.[48]

Based on our earlier finding that the point estimates are more positive for ineligible students, it is not surprising that the net benefits are far larger for this group. Our estimates of the net benefits for this subgroup are roughly double that of the eligible group under any set of assumptions.[49]

This analysis also addresses the general question of whether the results are robust to different approaches to the sampling weights. Two columns are shown for each eligibility category, one with trimmed weights and one with unadjusted weights. There is some movement in the net benefits when changing the weights, but the relative size of the net benefits between the eligible and ineligible groups is robust.

Even programs that pass a cost-benefit test might not be worthwhile, however, if there are other options that are more cost-effective (Harris, 2012, 2013). Such an analyses is challenging in this case because we are unaware of studies of other college access programs with strong identification strategies that also report both costs and outcomes for a similar range of education outcomes.

To test comparative cost-effectiveness, we adapted the analysis of Harris (2013) who calculated cost-effectiveness and benefit-cost ratios for a wide variety of programs where there is

[48] Some of the patterns in net benefits may be counter-intuitive. For example, in the first column, the net benefits decrease when we use larger returns to education. This occurs because some of the UB estimated effects are slightly negative (though insignificant), so a larger return actually reduces benefits. Also, the estimates in that same column are the same in the third and fifth rows. This is because all the MPR estimates are statistically insignificant and therefore set to zero in those rows, so -$7,301 reflects strictly the present discounted value of the direct costs of UB.

[49] We also carried out the analysis based on the income and first generation evidence in Table 8. As expected from the large coefficients for low-income-only, the benefits are larger for this group than for students who are both low-income and first generation. These results are omitted from Table 10 because of the previously noted limitation that we have no data on students who have higher incomes.

reasonably well-identified estimates of effectiveness.[50] Again, it is difficult to draw conclusions about effects on the average student. However, among typically ineligible students, UB appears more cost-effective than almost all the other options, including financial aid (with or without services), faculty-student ratios, full-time faculty, and general student services. Benefits have the potential to be larger for programs like UB that target student prior to college and, perhaps for this reason, the one program that does seem more cost-effective than UB is another TRIO program, Talent Search, although the effects of this program are not as well identified.[51] This is true regardless of what set of sampling weights we use.[52]

VI. Conclusion

It is now clear why the conclusions from the UB experiment about average treatment effects have been so controversial. The original study drew its main conclusions from its pre-specified design even though the implementation of that design apparently yielded biased estimates of the population average treatment effects. There are good reasons to stand by pre-specified analysis plans in general, but equally good reasons for expressing caution in stating conclusions when results from the original plan are not robust. It would therefore seem that the conclusions of both the original report and responses by critics are both too strong. The problems with the study design make it difficult to identify nationally representative average treatment effects for UB that are unbiased or precise.

The controversy over the average effects also drew attention away from what may be the more important policy question: who is being served by UB and how do the effects vary across

[50] The assumptions are quite similar to the earlier cost-benefit analysis, e.g., in terms of discounting and productivity growth.
[51] The program was studied using Propensity Score Matching (incidentally, that study was also carried out by MPR).
[52] In theory it might be possible to identify specific subgroups for which the other listed programs are also especially effective. Part of the point here is to show that this question is rarely considered.

these groups? For those students typically deemed ineligible, the social benefits are positive both in absolute terms and compared with many common alternatives, Even when site 69 is heavily weighted, the effects for this subgroup are positive and often significant for high school graduation, credential completion, and AA degrees. Compared with eligible students, almost every point estimate is larger for those who are typically ineligible. Similarly, the effects are larger for most outcomes in sites that do not use performance requirements to determine continued UB eligibility.

The bottom line is that UB would apparently be more beneficial if targeted to students who had more measureable behavioral difficulties—precisely the opposite direction of recent changes in policy and implementation.[53] Over time, the family income and education levels of UB students' parents have increased considerably, both in relative and absolute terms. Academically, UB students are now about average.

The drift in UB's focus away from its original mission of serving traditionally disadvantaged students was likely motivated by changes in how the federal government evaluates grantees. Program renewal decisions are now based on the percentage of students who enroll in and graduate from college, creating incentives for program operators to recruit and select students who are already likely to go to college regardless of program effectiveness. The easiest way to generate good college outcomes is to recruit and select students with a high likelihood of success, precisely those students who benefit least.[54]

Ironically, we might never have been able to test the hypothesis about student eligibility if not for one of the features of the experiment that has been often-criticized by critics of the

[53] See Harris (2014) for more on these conclusions and recommendations.
[54] This conclusion is not unlike that of other authors who have noted the fact that merit-based college programs redistribute benefits to the well off (e.g., Dynarski, 2000).

study: the over-subscription requirement (Cahalan, 2009). Experiments require sites to have many more applicants than participants so they can generate control and treatment groups. This may require some sites to attract a broader range of participants than occurs under business-as-usual operation, facilitating tests of the eligibility requirements themselves such as those carried out here. Policymakers should consider the value of over-subscription when debating whether to ban over-subscription as a requirement for study participation, as they have already done with TRIO programs.

The study also has a variety of implications for the design and analysis of large-scale randomized trials. Most obviously, it is best to avoid giving some observations 80 times more weight than others. The apparent mis-assignment of site 69 appears to have been coincidental, but the sampling design compounded the problem. If the sites had been more equally weighted, it would not have mattered as much that one site was wrongly placed.

With its lineage in the *Great Society*, Upward Bound is the nation's flagship college access programs and one of the oldest social mobility programs in the country. It will therefore always be a subject of political contention, especially when budgets are tight. With the best evidence to date, we find that the program probably passes a cost-benefit test as it currently operates, but could be twice as beneficial, and cost-effective compared with many alternatives, if it were better targeted. Getting back to the program's roots of serving highly disadvantaged students would improve college outcomes and simultaneously improve equity by addressing the growing gaps in college attainment between rich and poor.

References

Abadie, A., Chingos, M. & West, M. (2013). Endogenous Stratification in Randomized Experiments. *Unpublished working paper*. Harvard University.

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

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.

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

Bettinger, E.P. & Baker, R. (2014). The Effects of Student Coaching An Evaluation of a Randomized Experiment in Student Advising. *Educational Evaluation and Policy Analysis*, 36(1), pp. 3-19.

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.

Burkheimer, G. J., Levinsohn, Koo and French. (1976). *Evaluation Study of the Upward Bound Program: Volume IV*. Research Triangle Institute, Durham, NC.

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.

Castleman, B.L., Arnold, K.D., & Wartman, K.L. (2012). Stemming The Tide of Summer Melt: An Experimental Study of the Effects of Post-High School Summer Intervention on College Enrollment. *The Journal of Research on Educational Effectiveness*, 5(1): 1 – 18.

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*, 2651-2658.

Council on Opportunity in Education (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 MPR Fifth Follow Up Report), Prepared by MPR Policy Research*. Washington, DC: Council for Opportunity in Education.

Cox, B. G., and McGrath, D. S. (1981). An examination of the effect of sample weight truncation on the mean square error of estimates. presented at Biometrics Society ENAR meeting, Richmond VA.

Curtin, T.R., & Cahalan, M.W. (2004). *A Profile of the Upward Bound Math-Science Program: 2000-2001*. Washington, D.C.: U.S. Department of Education, Office of Postsecondary Education.

Decker, P. (2013). False Choices, Policy Framing, and the Promise of "Big Data." Presidential Address to the Association for Policy Analysis and Management. November 8, 2013. Washington, DC. Downloaded June 21, 2014 from: www.appam.org.

Dickens, W.T. (1990). Error Components in Grouped Data: Is It Ever Worth Weighting? *Review of Economics and Statistics*, 72(2): 328-33.

Dynarski, S. (2000). Hope for whom? Financial aid for the middle class and its impact on college attendance. *National Tax Journal*, 53(3), 629-661.

Field, K. (2007). Are the Right Students 'Upward Bound?' *Chronicle of Higher Education* 53(50), 16.

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

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.

Goldrick-Rab, S. & Harris, D. (2010). Observations on the use of NSC data for research purposes. Unpublished. Manuscript.

Greenleigh, A. (1970). *Upward Bound 1965-1969: A History and Synthesis of Data on the Program*. U.S. Office of Economic Opportunity. Washington, D.C.

Hansen, B.B. (2008). The Prognostic Analogue of the Propensity Score. *Biometrika* 95(2): 481-488.

Hanushek, E.A., Kain, J.F., O'Brien, D.M., Rivkin, S.G. (2005). The Market for Teacher Quality *NBER Working Paper No. 11154*. Cambridge, MA: National Bureau of Economic Research.

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

Harris, D. (2013). Applying cost-effectiveness analysis in higher education. In A. Kelly and K. Carey (eds.). *Stretching the Higher Education Dollar*. (pp. 45-66). Washington, DC: American Enterprise Institute.

Harris, D. (2014). Strengthening Federal Access Programs to Meet 21st Century Needs: A Look at TRIO and GEAR UP. Testimony to the U.S. Senate, Health, Education, Labor and Pensions Committee.

Harris, D. and Sass, T. (2011). Teacher training, teacher quality, and student achievement. *Journal of Public Economics*. 95: 798-812.

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.

Hoxby, C.M. (2000). Does Competition among Public Schools Benefit Students and Taxpayers? *American Economic Review*, 90(5), pp. 1209-1238.

Hoxby, C. & Turner, S. (2013). Expanding College Opportunities for High-Achieving, Low Income Students. *SIEPR Discussion Paper No. 12-014*. Palo Alto, CA: Stanford Institute for Economic Policy Research.

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

Jepsen, C., Troske, K., & Coomes P. (2014). The Labor-Market Returns to Community College Degrees, Diplomas, and Certificates. *Journal of Labor Economics*, 32(1), 95-121.

Johnson, N. (2009). *What does a college degree cost?* Washington, DC: Delta Cost Project.

Kane, T., and. C. Rouse. (1995): Labor market Returns to Two- and Four-Year College. *American Economic Review*, 85(3), 600-614.

Kolenikov , S. (2010). Resampling variance estimation for complex survey data. *The Stata Journal* 10(2), 165–199.

Krueger, A.B. & Zhu, P. (2003). Another Look at the New York City School Voucher Experiment, *NBER Working Paper No. 9418*. Cambridge, MA: National Bureau of Economic Research.

Levin, H., Belfield, C., Muennig, P., & Rouse, C. (2006). The Costs and Benefits of an Excellent Education for All of America's Children. Working Paper. Columbia University, Teachers College.

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.

Marcotte, D., Bailey, T., Borkoski, C., & Kienzl, G. (2005). The returns of a community college education: Evidence from the National Education Longitudinal Study. *Educational Evaluation and Policy Analysis*, 27(2), 157–175.

Mayer, D.P., Peterson, P.E., Myers, D.E., Tuttle, C., & Howell, W.G. (2002). *School Choice in New York City After Three Years: An Evaluation of the School Choice Scholarships Program*. Washington, DC: Mathematica Policy 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.

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. U.S. 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

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.

Potter, F. (1988). Survey of Procedures to Control Extreme Sampling Weights. *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.

Reardon, S.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.

Rothstein, J. (2007). Does Competition Among Public Schools Benefit Students and Taxpayers? A Comment on Hoxby (2000). *American Economic Review* 97(5), December 2007, pp. 2026-2037.

Sanbonmatsu, L., Kling, J.R., Duncan, G.J. & Brooks-Gunn, J. (2006). Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment. *Journal of Human Resources* 41(4): 649-691.

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

Solon, G., Haider, S.J., & Wooldridge, J. . Forthcoming. "What Are We Weighting For?" Journal of Human Resources.

Swail, W.S. (2001). Educational Opportunity 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

Tyler, J.H., Murnane, R.J., & Willett, J.B. (2000) Estimating the labor market signaling value of the GED. *Quarterly Journal of Economics*, 115(2), 431-468.

U.S. Department of Education (1988). U.S. Dept. of Education, National Center for Education Statistics. National Educational Longitudinal Study, 1988. Chicago, IL: National Opinion Research Center [producer], 1989. Ann Arbor, MI

U.S. Department of Education (2010). List of 2010 Upward Bound Grantees. Retrieved May 2, 20011 from: http://www2.ed.gov/programs/trioupbound/awards.html.

U.S. Department of Education (2012). Table 401. Revenues of public degree-granting institutions, by source of revenue and level of institution: Selected years, 2005-06 through 2010-11. Downloaded November 17, 2014 from: http://nces.ed.gov/programs/digest/d12/tables/dt12_401.asp.

U.S. Department of Education (2014). Table 330.10. Average undergraduate tuition and fees and room and board rates charged for full-time students in degree-granting postsecondary institutions, by level and control of institution: 1963-64 through 2012-13. Downloaded September 25, 2014 from: http://nces.ed.gov/programs/digest/d13/tables/dt13_330.10.asp.

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

This appendix includes discussion of key variables, provides additional formal analysis of the influence of weight errors in effect heterogeneity analysis, and includes additional tables and figures.

A1. Data

This section considers comparability of the NELS:88 and HSLS:09.

Survey Timing. NELS:88 data come from the survey's follow-up in 10th grade, while HSLS:09 data come from the 9th grade first follow-up survey, unless otherwise noted. Baseline characteristics from the experiment (MPR) are potentially measured at a variety of grades, so we restrict the respondent sample to students in the grade corresponding to the national survey in question when making comparisons.

Upward Bound Participation. In both national surveys we code students as having participated in UB if they indicated participation at any point. Information was not available to distinguish between participants in Upward Bound from Upward Bound Math and Science.

Students in NELS were only asked about UB participation if they were still enrolled in school as of the survey's second follow-up in 12th grade. We therefore restrict samples in all three datasets to students still enrolled at grade 12 for consistency in attrition requirements.

Race. Racial categories were collapsed from more detailed measures in the national surveys to align with the categories (Asian, Black, Hispanic, and White) in the MPR data. The Asian race category in MPR includes "Other" whereas the category is separate from "Other" in the national samples.

Language other than English at Home. In the NELS:88, the question in the survey is as follows: "Is any other language besides English spoken in your home?" In the HSLS:09, it is: "Is any language other than English regularly spoken in your home?"

Non-Native English Speaker. In the NELS:88, the question is: "Is English your native language?" asked in 10th grade survey. In the HSLS:09, 9th grade native language question suppressed in public-use file, so we created this variable from dual language question (non-native English speaker if any non-English first-language).

At Least one Parent Completed AA/BA+. A child is first generation if neither parent has BA or higher.

Family Income. For the mean family income category, the reported value is mean value of individual medians of income ranges in categorical response variables. The NELS:88 variable is adjusted to 2009 dollars. Distribution of income categories varies between datasets.

For the federal poverty line measure, the NELS:88 is the portion of the sample with income category below designated percentage of the federal poverty line as defined in 1988, adjusted for reported family size. The HSLS:09 contains direct measures of whether respondent's family was below 130% or 185% of the FPL. To create a measure comparable to the indicator for below 150% of the FPL in the MPR data, we take a weighted average of the means for 130% and 185%.

SES Quintile/Math Score Quintile. Each dataset contains a SES variable intended to have some degree of external comparability. We report means of quintile values to emphasize students' position in the wider distribution rather than the variable values. Math scores are presented as quintile means for the same reason.

Education Expectations. These measures are collapsed to be comparable between all three datasets. The table below documents how categories were combined.

NELS and HSLS Educational Expectations Measures

Category Reported in Paper Text	*NELS Categories*	*HSLS Categories*
Less than High School	Same	Same
High School	Same	High school diploma or GED
<2 Years of College/Vocational	≤ 2Years Trade, <2 Years College	Start an Associate's degree
2+ Years of College	Same	Complete AA, Start BA
BA	Finish college	Complete BA, Start MA
MA	Same	Complete MA, Start PhD/Prof.
PhD/Professional Degree	PhD, M.D.	Complete a PhD

Number of Times Parents Contacted About Behavior. In the NELS:88 (8th Grade Survey) the survey question is: "Since your eighth grader's school opened last Fall, how many times HAVE YOU BEEN CONTACTED BY THE SCHOOL about the following? Your eighth grader's behavior in school" Categorical responses coded as follows: "NONE" (0), "ONCE OR TWICE" (1.5), "THREE OR FOUR TIMES" (3.5), "MORE THAN FOUR TIMES" (7). In the HSLS, the question is: "During the last school year (2008-2009), how many times were you or another family member contacted by the school about [your 9th grader]'s problem behavior in school?" Categorical responses coded identically to NELS.

Ever Suspended or on Probation. In the NELS:88, there is an indicator for whether parent responded that student had ever been suspended or expelled as of 10th grade survey. In the HSLS:09, the indicator is for whether the parent responded that student had ever been suspended or expelled as of 9th grade survey. In the experimental sample, the indicator is for whether student had ever been suspended or on probation by grade at baseline.

A2: Role of Weight Errors in Effect Heterogeneity Analysis

In section III, we discussed, how errors in the weights influence may introduce bias in the treatment effects. To extend this to the effect heterogeneity analysis, we extend equation (4) so that there are two subgroups indexed by 1 and 2 (and site types continue to be indexed by letters A and B). We estimate and report separate effects for the two subgroups:

$$\hat{\beta}_1 = (\widehat{w}_{1A}\beta^*_{1A} + \widehat{w}_{1B}\beta^*_{1B}) = \beta^*_1 + (e_{1A}\beta^*_{1A} + e_{1B}\beta^*_{1B}) \quad (5a)$$

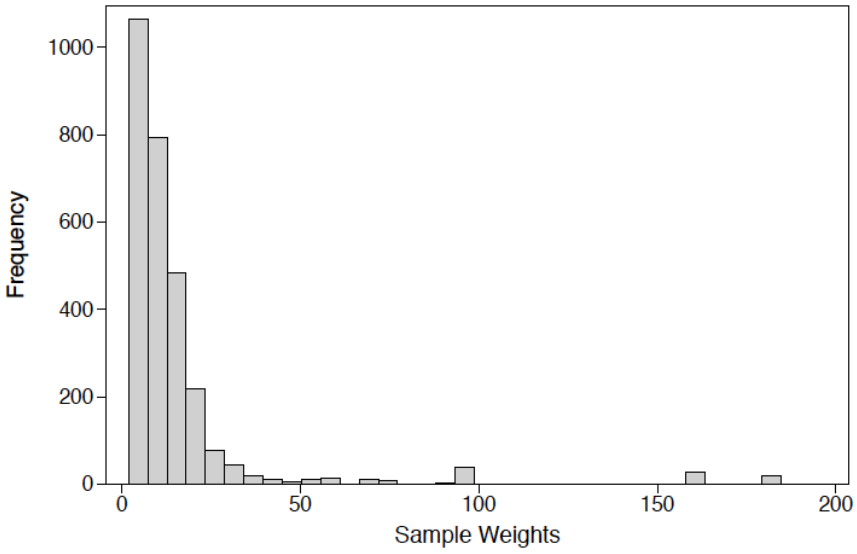
$$\hat{\beta}_2 = (\widehat{w}_{2A}\beta^*_{2A} + \widehat{w}_{2B}\beta^*_{1B}) = \beta^*_2 + (e_{2A}\beta^*_{2A} + e_{2B}\beta^*_{2B}) \quad (5b)$$

The difference in reported effects ($\hat{\beta}_1 - \hat{\beta}_2$) is unbiased when $E[(e_{1A}\beta^*_{1A} + e_{1B}\beta^*_{1B}) - (e_{2A}\beta^*_{2A} + e_{2B}\beta^*_{2B})] = 0$. With subgroup analysis, the condition can hold even when the covariance between the errors and effect estimates is non-zero for every site-by-group. If $\mu_e = 0$ within each subgroup-by-site (as before), then the condition becomes: $Cov(e_{1A}, \beta^*_{1A}) + Cov(e_{1B}, \beta^*_{1B}) - Cov(e_{2A}, \beta^*_{2A}) - Cov(e_{2B,}\beta^*_{2B}) = 0$. This is more plausible than the condition for the average treatment effects. For example, the covariances might be non-zero because type A sites are less organized, leading to both worse data for assigning strata and creating the weights, and lower program efficiency (across subgroups). But it is also possible that the covariance are of the opposite sign so that the problem worsens. Unfortunately, the assumption is not testable.

A3: Miscellaneous Tables and Figures

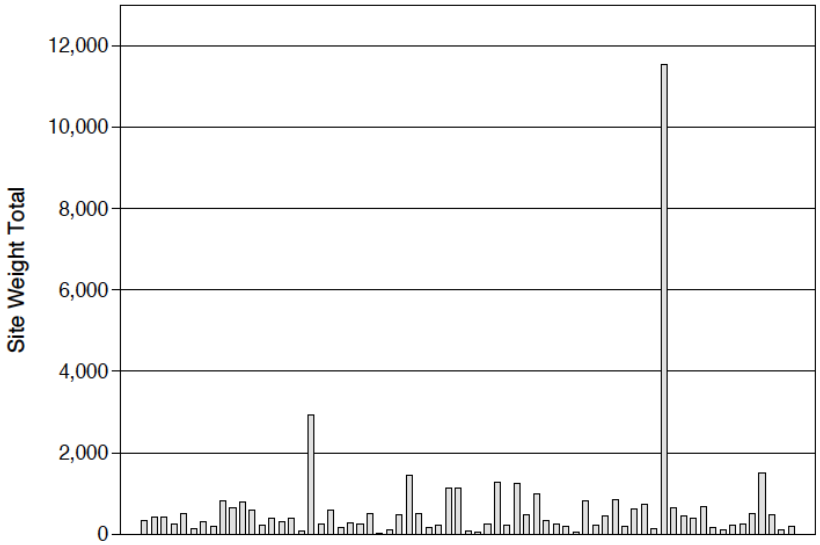
The following tables and figures are referenced in the main text.

Figure A1: Sampling Weights Histogram



Notes: Figure A1 includes the number of observations (y-axis) that have various weights, up to 180.

Figure A2: Weighted Population Totals, By Site



Notes: Figure A2 indicates the total individual weights by site. Site 69 is the outlier toward the right.

Table A1: Comparison of Harris, Nathan, and Marksteiner (HNM) and MPR Variable Averages

Variable	*MPR Published Results*	*HNM PPPS data files*
Low-income household	0.85	0.85
At least one parent has a B.A.	0.06	0.06
Female	0.68	0.67
White	0.28	0.28
Hispanic	0.19	0.19
Black	0.43	0.43
Other race	0.10	0.10
≤9th grade at application	0.54	0.54
10th grade or above	0.46	0.46

Notes: Figures are unweighted. "HMN" are based on authors' calculations. The MPR data from table II.3 of their final report.

Appendix A2: Screening Measures for Eligibility

Student Survey Question	*Threshold Response*	*Response Range*	*% Ineligible*
During the last school year, how often did each of the following things happen to you?			
I was late for school	10+ times	0-10+ times	5.41
I cut or skipped classes	10+ times	0-10+ times	1.78
I missed a day of school	10+ times	0-10+ times	10.29
I got in trouble for not following school rules	10+ times	0-10+ times	1.74
I was put on in-school suspension	3+ times	0-10+ times	2.49
I was suspended or put on probation from school	1+ times	0-10+ times	9.96
I was transferred to another school for disciplinary reasons	1+ times	0-10+ times	1.03
I was arrested	1+ times	0-10+ times	1.88
I spent time in a juvenile home/detention center	1+ times	0-10+ times	0.96
How sure are you that you will graduate from high school?	I probably won't graduate	Probably won't-Definitely will	0.57
How far in school do you think you will go?	< 2 years of college	Won't finish high school-Earn PhD	9.20

Notes: The questions are direct quotations from the surveys administered to grantees. The % ineligible is based on average eligibility criteria. We also use site-by-site criteria elsewhere in the paper.

Table A3: Effects by Predicted Outcomes

	Unweighted		*Weighted*		Group Mean
	Est.	S.E.	Est.	S.E.	Outcome
BA Degree +					
Low	0.000	0.045	0.004	0.023	0.135
Med	0.068	0.040	0.065	0.034	0.140
High	0.045	0.289	0.033	0.041	0.279
AA Degree +					
Low	0.024	0.068	0.034	0.032	0.193
Med	0.015	0.046	-0.001	0.045	0.293
High	0.034	0.339	0.032	0.042	0.357
Certificate +					
Low	0.061	0.202	0.050	0.030	0.332
Med	0.053	0.061	0.051	0.048	0.354
High	0.070	0.277	0.046	0.041	0.453
College Enrollment					
Low	0.037	0.022	0.008	0.038	0.780
Med	0.054	0.034	0.038	0.031	0.825
High	0.028	0.049	0.038	0.025	0.865
High School GPA					
Low	0.080	0.126	0.065	0.051	1.977
Med	-0.028	0.066	0.006	0.062	2.444
High	0.067	0.310	0.084	0.048	2.771
High School Grad.					
Low	0.041	0.055	0.041	0.030	0.746
Med	0.070	0.033	0.047	0.024	0.865
High	0.029	0.025	0.029	0.019	0.928

Notes: Results based on LOO-SW method described in text with 500 iterations. Means column reports actual outcome means for the control group. All results presented here use 30/75 weighted second stage with full MPR covariates list. Notes: Results based on modified Leave One Out (LOO) method. The first stage prediction model is unweighted and includes discipline incidents, demographics, and site-specific intercepts as covariates. Second stage is the same as column (6) in Table 3.

Figure 1: Site Effect Heterogeneity

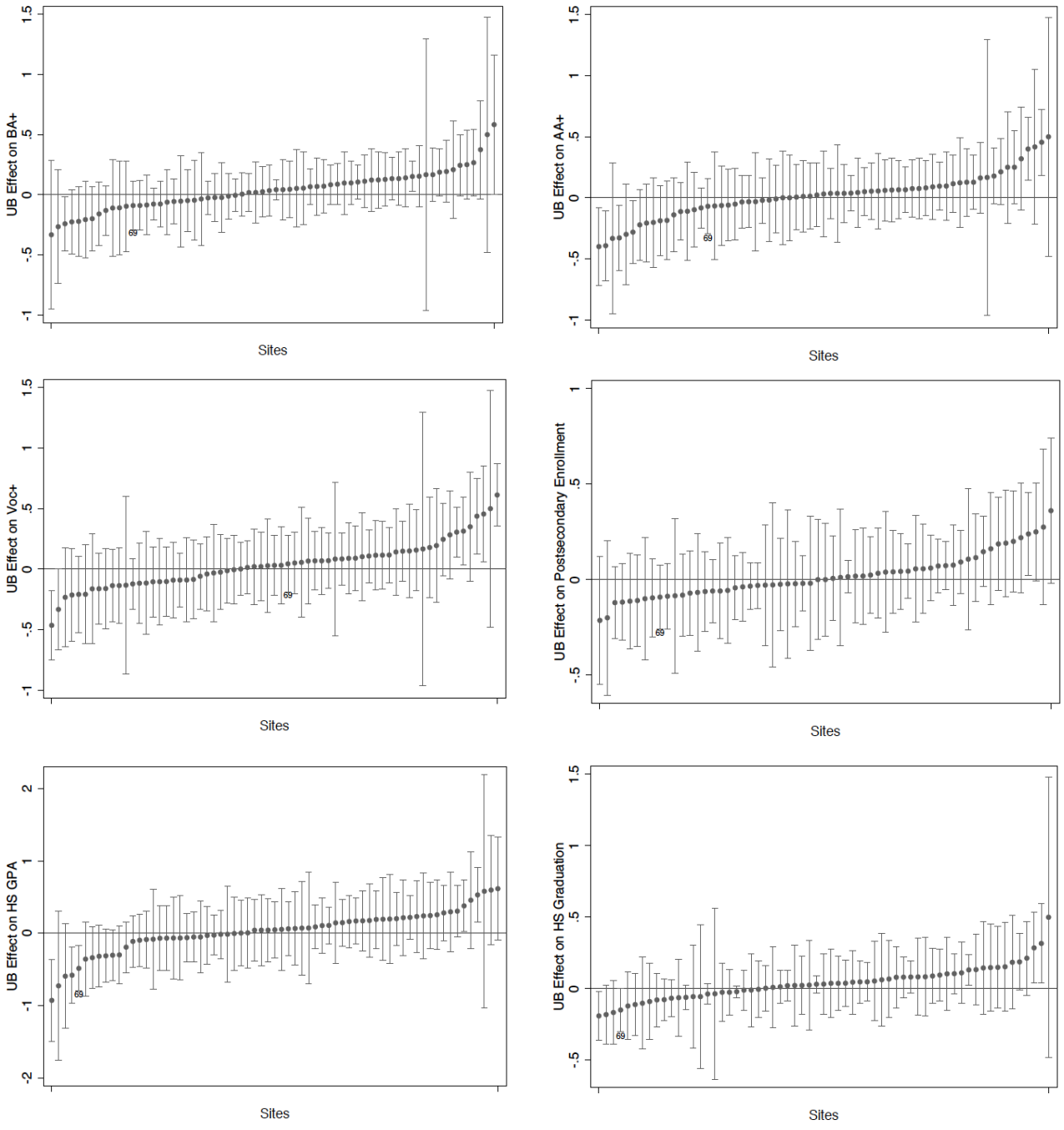


Table 1: Descriptive Statistics for Covariates

Variable	*Obs*	*Mean*	*Std. Dev.*	*Min*	*Max*
Female	2844	0.670	0.470	0	1
White	2844	0.277	0.448	0	1
Hispanic	2844	0.186	0.389	0	1
Black	2844	0.434	0.496	0	1
Other Race	2844	0.104	0.305	0	1
Applied to UB in 8th grade	2844	0.147	0.354	0	1
Applied to UB In 9th grade	2844	0.392	0.488	0	1
Applied to UB in 10th grade	2844	0.355	0.479	0	1
Applied to UB in 11th grade	2844	0.107	0.309	0	1
First generation in college (only)	2844	0.149	0.357	0	1
Low income household (only)	2844	0.055	0.228	0	1
Student expects PhD	2844	0.250	0.433	0	1
Student expects MA	2844	0.133	0.340	0	1
Student expects BA	2844	0.355	0.479	0	1
Student expects AA	2844	0.117	0.321	0	1
Student expects < 2 years college	2844	0.036	0.187	0	1
Student expects h.s. degree	2844	0.023	0.148	0	1
"Don't Know" on ed. expectations	2844	0.077	0.267	0	1
Flag for Site 69	2844	0.030	0.228	0	1

Notes: Estimates are unweighted. For simplicity, we use the same covariates and coding as Mathematica and these are shown above. The regression results shown later are also from models that include interactions between the Flag for Site 69 and all the other covariates to match MPR. All subsequent tables apply survey weights, except where noted.

Table 2: Baseline Equivalence

Variables	*Treatment*	*Control*	*Difference*	*Sig.*
Female	0.713	0.676	-0.037	
White	0.203	0.219	0.016	
Black	0.509	0.502	-0.007	
Hispanic	0.225	0.210	-0.015	
Other Race	0.063	0.068	0.005	
Applied to UB in 8th Grade	0.131	0.126	-0.005	
Applied to UB in 9th Grade	0.450	0.477	0.027	
Applied to UB in 10th Grade	0.329	0.300	-0.029	
Applied to UB in 11th Grade	0.090	0.096	0.007	
First-Generation College (only)	0.166	0.171	0.005	
Low-Income Household (only)	0.048	0.041	-0.006	
Student Expects PhD	0.272	0.215	-0.057	
Student Expects MA	0.153	0.095	-0.058	
Student Expects BA	0.328	0.373	0.045	
Student Expects AA	0.107	0.134	0.027	**
Student Expects <2 Years College	0.034	0.064	0.032	*
Student Expects HS Diploma	0.029	0.027	-0.002	
Reported "Don't Know" Educ. Expect.	0.077	0.090	0.013	
Site 69 Dummy	0.264	0.264	-2.26E-09	
F-statistics for overall differences				
Unweighted	27.16			
Sampling Weights (SW)	4.29			
SW Trimmed at 30/75 pct	84.44			

Notes: Sampling weights are applied therefore the control and treatment means differ somewhat from Table 1. * $p < .05$, ** $p < .01$, *** $p < .001$. Results are very similar to MPR; see their Table A.5 in the 2009 report and Table B.1 in the 2004 report.

Table 3: Main Impacts of Upward Bound

Dependent Variable	*Control Mean*	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
BA Degree+	0.186	0.0248	0.0349*	0.0217	0.0204	0.0401**	0.0360*
N [s.e.]	2,659	[0.0149]	[0.0134]	[0.0235]	[0.0233]	[0.0136]	[0.0152]
AA Degree+	0.282	0.0124	0.0212	0.0114	0.00987	0.0261	0.0114
	2,659	[0.0167]	[0.0160]	[0.0261]	[0.0257]	[0.0166]	[0.0261]
Certificate+	0.382	0.0342	0.0417	0.0545**	0.0514**	0.0502*	0.0545**
	2,659	[0.0218]	[0.0207]	[0.0175]	[0.0178]	[0.0212]	[0.0175]
College Enrollment	0.824	0.0182	0.0253*	0.014	0.00854	0.0268*	0.0253*
	2,659	[0.0133]	[0.0118]	[0.0314]	[0.0282]	[0.0131]	[0.0118]
High School GPA	2.395	0.0444	0.053	-0.00546	-0.0135	0.0461	0.053
	2,674	[0.0342]	[0.0317]	[0.0816]	[0.0779]	[0.0317]	[0.0317]
High School Grad	0.845	0.0342*	0.0388**	0.0297	0.0274	0.0419**	0.0388**
	2,649	[0.0146]	[0.0130]	[0.0329]	[0.0310]	[0.0144]	[0.0130]
Covariates			Y	Y	Y	Y	Y
Weighting Options							
No Weights	Y	Y	Y				
Sampling Weights				Y	Y		
SW Trimmed (AM)						Y	
SW Trimmed (MSE)							Y
Site-Spec. Intercepts					Y		

Notes: Standard errors in parentheses (clustered at the site level). All estimates are intent to treat (ITT). "AA Degree +" means AA degree or above. "Certification +" means any credential. In model (5), exclusion of site fixed effects is easily rejected for all outcomes. The SW Trimmed weights are based on the method proposed by Asparouhov and Muthen (2005) with trimming at the 30th and 75th percentiles. Standard errors are in parentheses. * $p < .05$, ** $p < .01$, *** $p < .001$.

Table 4: Comparison of Main Impacts with MPR Preferred

Dependent Variable	*(A) MPR Preferred*	*(B) MPR Other*	*(1) HNM Replicate Other*	*(2)*	*(3)*	*(4) HNM Preferred*
BA Degree	0.0014	0.0425	0.0388	0.0456*	0.0217	0.0360*
	n.r.	n.r.	[0.0220]	[0.0222]	[0.0235]	[0.0152]
N	n.r.	n.r.	1724	1724	2659	2659
AA Degree+	-0.0218	0.0356	0.0328	0.0367	0.0114	0.0114
	n.r.	n.r.	[0.0185]	[0.0202]	[0.0261]	[0.0261]
N	n.r.	n.r.	1724	1724	2659	2659
Certificate +	0.0454	0.1116	0.1161*	0.1162*	0.0545**	0.0545**
	n.r.	n.r.	[0.0563]	[0.0463]	[0.0175]	[0.0175]
N	n.r.	n.r.	1724	1724	2659	2659
College Enroll (Any)	0.0154	0.0156	0.0133	0.0136	0.014	0.0253*
	n.r.	n.r.	[0.0336]	[0.0334]	[0.0314]	[0.0118]
N	n.r.	n.r.	2102	2102	2659	2659
High School GPA	0	0	-0.0190	-0.0055	-0.0055	0.0530
	n.r.	n.r.	[0.1070]	[0.0816]	[0.0816]	[0.0317]
N	n.r.	n.r.	2006	2674	2674	2674
High School Grad.	-0.01	-0.01	-0.0010	0.0099	0.0297	0.0388**
	n.r.	n.r.	[0.0390]	[0.0364]	[0.0329]	[0.0130]
N	n.r.	n.r.	1895	2291	2649	2649
Weighting Options						
Non-Resp. Weights	Y	Y	Y			
Sampling Weights (SW)				Y	Y	
SW Trimmed (MSE)						Y
Data Sources						
High school: 3rd survey only	Y	Y	Y	Y		
College: 5th survey, NSC & SFA	Y					
College: 5th survey only		Y	Y	Y		
All Survey Waves (no NSC, SFA)					Y	Y

Notes: Columns (A) and (B) are copied from the MPR reports. Column (A) cannot be replicated due to data limitations. Column (1) is our replication of (B). Estimates from the 2004 reports are rounded in the MPR report as shown. Column (4) is the same as Column (6) in Table 3 and shown for comparison purposes. The GPA results for columns (2) and (3) are identical because this is the only outcome variable that comes from the transcripts, so the survey waves are not relevant. All estimates are covariate-adjusted. Standard errors are clustered at the site level. * p < .05, ** p < .01, *** p < .001.

Table 5A: Tests of Representativeness of UB Experimental Sample
(UB-only samples)

	NELS: 1988	Experiment Sample	HSLS: 2009	NELS – Experiment	HSLS – Experiment
	(1)	(2)	(3)	(1) – (2)	(3) – (2)
Female	0.506	0.715	0.444	-0.209***	-0.271***
	(0.036)	(0.033)	(0.028)		
Black	0.424	0.478	0.405	-0.054	-0.073
	(0.042)	(0.043)	(0.038)		
Hispanic	0.178	0.230	0.178	-0.051	-0.052
	(0.031)	(0.079)	(0.028)		
White	0.342	0.226	0.354	0.116	0.127
	(0.042)	(0.083)	(0.029)		
Non-Native English Speaker	0.135	0.100	0.131	0.035	0.031
	(0.034)	(0.023)	(0.031)		
At least one Parent Completed AA+	0.146	--	0.420		--
	(0.026)	--	(0.032)		
At least one Parent Completed BA+	0.107	0.039	0.272	0.068*	0.233***
	(0.023)	(0.014)	(0.029)		
Mean Family Income Cat. Median (2009 $)[1]	46.311	--	51.865	--	--
	(3.136)	--	(2.727)		
SES Quintile[2]	2.267	--	2.647	--	--
	(0.089)	--	(0.08)		
Family Under 130% Poverty Line[3]	0.435	--	0.394	--	--
	(0.038)	--	(0.037)		
Family Under 150% Poverty Line[4]	0.483	0.814	0.477	-0.331***	-0.337***
	(0.039)	(0.03)	(0.035)		
Family Under 185% Poverty Line	0.531	--	0.524	--	--
	(0.039)	--	(0.033)		
Under 130% Poverty & First Generation[5]	0.409	--	0.361	--	
	(0.038)	--	(0.038)		
Under 150% Poverty & First Generation	0.458	0.775	0.431	-0.316***	-0.344***
	(0.039)	(0.021)	(0.036)		
Under 185% Poverty & First Generation	0.502	--	0.471	--	
	(0.04)	--	(0.036)		
Took Algebra or Higher in 9th Grade	0.629	0.631	0.749	-0.002	0.118*
	(0.054)	(0.049)	(0.027)		
Math Score Quintile (1-5)	2.523	--	2.419	--	--
	(0.101)	--	(0.087)		
12th Grade Cumulative GPA	2.448	2.384	--	0.064	--
	(0.08)	(0.076)	--		

Table 5B: Tests of Representativeness of UB Experimental Sample

(UB-only samples)

	NELS: 1988	Experiment Sample		HSLS: 2009	NELS - Experiment	HSLS - Experiment
	10th (1)	9th (2)	10th (3)	9th (4)	10th (1)-(3)	9th (4)-(2)
Education Expectations						
Less than High School	0.003	0.027	0.004	0.026	-0.001	-0.001
	(0.003)	(0.018)	(0.003)	(0.01)		
High School	0.032	0.016	0.013	0.162	0.019	0.146***
	(0.012)	(0.007)	(0.006)	(0.021)		
<2 Years of College/Voc.	0.115	0.077	0.065	0.022	0.050	-0.054**
	(0.028)	(0.016)	(0.022)	(0.01)		
2+ Years of College	0.155	0.078	0.136	0.089	0.020	0.010
	(0.049)	(0.009)	(0.016)	(0.016)		
BA	0.326	0.435	0.463	0.184	-0.137	-0.251***
	(0.047)	(0.059)	(0.062)	(0.021)		
MA	0.202	0.111	0.103	0.228	0.098	0.117*
	(0.048)	(0.04)	(0.033)	(0.023)		
PhD/Prof. Degree	0.166	0.257	0.215	0.289	-0.048	0.032
	(0.032)	(0.038)	(0.019)	(0.031)		
Times Parents Contacted about Behavior[6]	0.984	--	--	0.991	--	--
	(0.117)	--	--	(0.148)		
Ever Suspended or on Probation[7]	0.231	0.394	0.327	0.226	-0.163	-0.167
	(0.047)	(0.205)	(0.187)	(0.029)		
Overall N	253	1,275		905		

Notes: The top panels (5A) are based on data from grades 11 and 12, while the bottom panels focus on grades closer to the pre-treatment period. As discussed in the text the grade of data collection does not perfectly align across the data sets. In NELS and HSLS, UB participation is defined as having ever noted program participation. In the MPR data, participation is defined based on assignment, although the results are very similar when defining based on UB participation. *** p<.001, ** p<.01, * p<.05, + p<.10

[1] Income categories converted to medians based on category endpoints and displayed mean.

[2] SES based on five equally weighted, standardized components: father's education, mother's education, family income, father's occupation, and mother's occupation.

[3] NELS federal poverty line estimates imputed from income categories; HSLS is actual reported.

[4] HSLS 150% poverty line values are a linear imputation from the 130% and 185% values; NELS is actual reported.

[5] First generation defined as neither parent has BA+.

[6] NELS parent behavior contact question comes from 8th grade data; HSLS is 9th grade.

[7] HSLS variable is "Ever Suspended or Expelled." NELS variable includes probation, as indicated in the left column.

Table 6A: Effects on Typically Eligible Students

Dependent Variable	(1)	(2)	(3)	(4)	(5)	(6)
BA Degree+	0.0223	0.0330*	0.0509**	0.0432*	0.0498**	0.0478**
	[0.0168]	[0.0145]	[0.0146]	[0.0162]	[0.0161]	[0.0139]
N	2033	2033	2033	2033	2223	2398
AA Degree+	-0.00681	0.00233	0.0131	0.0113	0.0276	0.0223
	[0.0175]	[0.0161]	[0.0192]	[0.0174]	[0.0182]	[0.0164]
N	2033	2033	2033	2033	2223	2398
Certificate+	0.0129	0.0213	0.0187	0.0271	0.0415	0.0412
	[0.0219]	[0.0199]	[0.0280]	[0.0214]	[0.0217]	[0.0217]
N	2033	2033	2033	2033	2223	2398
College Enrollment	0.0202	0.0258*	0.00609	0.0253	0.0290*	0.0326*
	[0.0130]	[0.0124]	[0.0362]	[0.0129]	[0.0133]	[0.0158]
N	2033	2033	2033	2033	2223	2398
High School GPA	0.0402	0.0406	0.0162	0.041	0.0354	0.0424
	[0.0365]	[0.0354]	[0.0736]	[0.0362]	[0.0311]	[0.0310]
N	2050	2050	2050	2050	2236	2417
High School Graduation	0.0275*	0.0298*	0.0112	0.0328*	0.0341*	0.0374*
	[0.0133]	[0.0123]	[0.0316]	[0.0137]	[0.0142]	[0.0142]
N	2019	2019	2019	2019	2221	2392
Covariates		Y	Y	Y	Y	Y
Weights						
No weights	Y	Y				
Sampling Weights (SW)			Y			
SW Trim (AM)				Y	Y	Y
Identifying At Risk						
Avg. Site w/o expectations	Y	Y	Y	Y		
By Site w/ expectations					Y	
By Site w/o expectations						Y

Notes: Model specifications are identical to Table 3; Standard errors are in parentheses (clustered at site level). * p < .05, ** p < .01, *** p < .001.

Table 6B: Effects on Typically Ineligible Students

Dependent Variable	*(1)*	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*
BA Degree+	0.0343	0.0386	-0.00552	0.0323	-0.00142	-0.015
	[0.0238]	[0.0243]	[0.0365]	[0.0242]	[0.0215]	[0.0347]
N	626	626	626	626	436	261
AA Degree+	0.0761*	0.0764*	0.0287	0.0695*	0.0217	0.064
	[0.0311]	[0.0294]	[0.0485]	[0.0274]	[0.0337]	[0.0508]
N	626	626	626	626	436	261
Certificate+	0.104*	0.101*	0.110*	0.117**	0.0648	0.103
	[0.0453]	[0.0446]	[0.0429]	[0.0426]	[0.0454]	[0.0645]
N	626	626	626	626	436	261
College Enrollment	0.0126	0.0257	0.0535	0.0351	0.0053	-0.0231
	[0.0361]	[0.0358]	[0.0412]	[0.0364]	[0.0427]	[0.0502]
N	626	626	626	626	436	261
High School GPA	0.0672	0.0855	0.0035	0.0716	0.107	0.0935
	[0.0659]	[0.0617]	[0.0959]	[0.0654]	[0.0802]	[0.0912]
N	624	624	624	624	438	257
High School Grad	0.0576	0.0696	0.109*	0.0777*	0.0734	0.0814
	[0.0336]	[0.0349]	[0.0462]	[0.0380]	[0.0603]	[0.0541]
N	630	630	630	630	428	257
Covariates		Y	Y	Y	Y	Y
Weights						
No weights	Y	Y				
Sampling Weights (SW)			Y			
SW Trim (AM)				Y	Y	Y
Identifying At Risk						
Avg. Site w/o expectations	Y	Y	Y	Y		
By Site w/ expectations					Y	
By Site w/o expectations						Y
P-values of differences						
BA Degree+	0.657	0.835	0.111	0.699	0.080	0.090
AA Degree+	0.013	0.017	0.705	0.041	0.876	0.412
Certificate+	0.050	0.072	0.094	0.035	0.590	0.331
College Enrollment	0.843	0.998	0.326	0.804	0.605	0.332
High School GPA	0.718	0.540	0.865	0.682	0.418	0.611
High School Graduation	0.372	0.272	0.037	0.258	0.542	0.417

Notes: The *p*-values for differences come from the coefficient on the interaction between eligibility status and treatment assignment in a regression analysis that includes the entire sample. Standard errors are in parentheses (clustered at site level). * p < .05, ** p < .01, *** p < .001.

Table 7A: National Trends in UB Population Family Background

	NELS: 1988			HSLS: 2009			
	Upward Bound	Non-UB	Diff	Upward Bound	Non-UB	Diff	Diff-in-Diff
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	(7)
Female	0.506	0.506	0.000	0.444	0.507	0.062	0.063
	(0.036)	(0.005)	(0.036)	(0.028)	(0.007)	(0.028)	(0.046)
Asian	0.022	0.034	0.012	0.054	0.043	-0.011	-0.023
	(0.009)	(0.002)	(0.009)	(0.013)	(0.005)	(0.014)	(0.016)
Black	0.424	0.097	-0.327	0.405	0.162	-0.243	0.084
	(0.042)	(0.006)	(0.041)	(0.038)	(0.01)	(0.035)	(0.054)
Hispanic	0.178	0.089	-0.089	0.178	0.203	0.026	0.115**
	(0.031)	(0.007)	(0.03)	(0.028)	(0.009)	(0.027)	(0.04)
White	0.342	0.761	0.419	0.354	0.585	0.231	-0.188***
	(0.042)	(0.01)	(0.041)	(0.029)	(0.013)	(0.028)	(0.049)
Lang. other than English at Home	0.264	0.160	-0.104	0.228	0.230	0.003	0.106*
	(0.04)	(0.008)	(0.039)	(0.037)	(0.009)	(0.036)	(0.054)
Non-Native English Speaker	0.135	0.079	-0.055	0.131	0.113	-0.018	0.038
	(0.034)	(0.006)	(0.034)	(0.031)	(0.006)	(0.028)	(0.044)
Attends Urban School	0.468	0.236	-0.231	0.389	0.308	-0.080	0.151*
	(0.057)	(0.014)	(0.056)	(0.041)	(0.004)	(0.043)	(0.071)
At least one Parent Completed AA+	0.146	0.343	0.197	0.420	0.550	0.130	-0.067
	(0.026)	(0.008)	(0.027)	(0.032)	(0.012)	(0.035)	(0.044)
At least one Parent Completed BA+	0.107	0.296	0.189	0.272	0.388	0.116	-0.073+
	(0.023)	(0.008)	(0.024)	(0.029)	(0.012)	(0.031)	(0.039)
Mean Family Inc. Cat. Median (2009 $)[1]	46.311	74.129	27.818	51.865	72.724	20.859	-6.959*
	(3.136)	(1.163)	(1.776)	(2.727)	(1.533)	(2.984)	(3.473)
SES Quintile[2]	2.267	3.063	0.796	2.647	3.092	0.445	-0.351**
	(0.089)	(0.027)	(0.092)	(0.08)	(0.036)	(0.084)	(0.125)
Family Under 130% Poverty Line[3]	0.435	0.181	-0.254	0.394	0.239	-0.156	0.098+
	(0.038)	(0.006)	(0.038)	(0.037)	(0.011)	(0.036)	(0.052)
Family Under 150% Poverty Line[4]	0.483	0.231	-0.252	0.477	0.321	-0.155	0.097+
	(0.039)	(0.007)	(0.039)	(0.035)	(0.012)	(0.037)	(0.051)
Family Under 185% Poverty Line	0.531	0.275	-0.257	0.524	0.369	-0.155	0.102*
	(0.039)	(0.008)	(0.039)	(0.033)	(0.013)	(0.032)	(0.054)
Under 130% Poverty & First Generation[5]	0.409	0.167	-0.242	0.361	0.214	-0.146	0.095+
	(0.038)	(0.006)	(0.038)	(0.038)	(0.011)	(0.038)	(0.054)
Under 150% Poverty & First Generation	0.458	0.212	-0.247	0.431	0.284	-0.147	0.099
	(0.039)	(0.007)	(0.04)	(0.036)	(0.477)	(0.478)	(0.48)
Under 185% Poverty & First Generation	0.502	0.251	-0.251	0.471	0.323	-0.147	0.104+
	(0.04)	(0.007)	(0.04)	(0.036)	(0.013)	(0.037)	(0.054)

Table 7B: National Trends in UB Population Academic Background

	NELS: 1988			HSLS: 2009			
	UB	Non-UB	Diff	UB	Non-UB	Diff	Diff-in-Diff
	(1)	*(2)*	*(3)*	*(4)*	*(5)*	*(6)*	(7)
Took Algebra or Higher in 9th Grade	0.629	0.695	0.066	0.749	0.842	0.093	0.027
	(0.054)	(0.008)	(0.055)	(0.027)	(0.008)	(0.027)	(0.061)
Math Score Quintile (1-5)	2.523	3.142	0.619	2.419	3.106	0.688	0.069
	(0.101)	(0.022)	(0.101)	(0.087)	(0.034)	(0.094)	(0.138)
Education Expectations							
Less than High School	0.003	0.004	0.000	0.026	0.005	-0.021	-0.022*
	(0.003)	(0.001)	(0.003)	(0.01)	(0.001)	(0.01)	(0.011)
High School	0.032	0.081	0.049	0.162	0.167	0.005	-0.044+
	(0.012)	(0.005)	(0.013)	(0.021)	(0.007)	(0.023)	(0.027)
<2 Years of College/Vocational	0.115	0.145	0.030	0.022	0.008	-0.015	-0.044
	(0.028)	(0.005)	(0.029)	(0.01)	(0.001)	(0.01)	(0.03)
2+ Years of College	0.155	0.143	-0.012	0.089	0.084	-0.005	0.007
	(0.049)	(0.005)	(0.05)	(0.016)	(0.004)	(0.016)	(0.052)
BA	0.326	0.336	0.010	0.184	0.228	0.044	0.034
	(0.047)	(0.006)	(0.048)	(0.021)	(0.006)	(0.021)	(0.052)
MA	0.202	0.155	-0.047	0.228	0.264	0.035	0.082
	(0.048)	(0.005)	(0.048)	(0.023)	(0.006)	(0.025)	(0.054)
PhD/Professional Degree	0.166	0.136	-0.031	0.289	0.245	-0.044	-0.013
	(0.032)	(0.005)	(0.032)	(0.031)	(0.006)	(0.033)	(0.046)
Times Parents Contacted about Behavior[6]	0.984	0.615	-0.369	0.991	0.584	-0.408	-0.0390
	(0.117)	(0.015)	(0.118)	(0.148)	(0.021)	(0.147)	(0.188)
Ever Suspended or on Probation[7]	0.231	0.129	-0.101	0.226	0.125	-0.101	0.0000
	(.047)	(.006)	(.048)	(0.029)	(0.007)	(0.029)	(0.056)

Notes: The notes from Table 5B for variable definitions also apply here. The difference-in-difference column is calculated as follows: (Non-UB09 – UB09) – (Non-UB88 - UB88). The HSLS results also omit participants in the federal Talent Search program. See other important notes in the text explaining these figures. Statistical significance tests are only conducted on the DD coefficients: *** p<.001, ** p<.01, * p<.05, + p<.10.

Table 8: Effect Heterogeneity by Low-Income/First-Generation Status

	(2)	*(3)*	*(4)*	*(5)*	*(6)*
BA Degree+					
UB*Both	0.0291*	0.00632	0.00521	0.0346*	0.0281
	[0.0142]	[0.0324]	[0.0322]	[0.0147]	[0.0178]
UB*First-Gen Only	-0.0125	0.0511	0.0567	-0.0127	0.00545
	[0.0458]	[0.0643]	[0.0619]	[0.0517]	[0.0532]
UB*Low-Inc Only	0.134*	0.145	0.122	0.135	0.133
	[0.0618]	[0.0875]	[0.0862]	[0.0812]	[0.0847]
AA Degree+					
UB*Both	0.0172	-0.00982	-0.0115	0.0200	-0.00982
	[0.0164]	[0.0388]	[0.0386]	[0.0181]	[0.0388]
UB*First-Gen Only	-0.0127	0.0786	0.0861	-0.0087	0.0786
	[0.0441]	[0.0841]	[0.0830]	[0.0471]	[0.0841]
UB*Low-Inc Only	0.105	0.174*	0.149	0.137	0.174*
	[0.0823]	[0.0855]	[0.0822]	[0.0829]	[0.0855]
Certificate+					
UB*Both	0.0387	0.0296	0.0258	0.0444	0.0296
	[0.0214]	[0.0270]	[0.0271]	[0.0238]	[0.0270]
UB*First-Gen Only	0.0208	0.137	0.148	0.0335	0.137
	[0.0395]	[0.0932]	[0.0906]	[0.0430]	[0.0932]
UB*Low-Inc Only	-0.0033	0.0488	0.0209	0.0112	0.0488
	[0.0782]	[0.0680]	[0.0620]	[0.0752]	[0.0680]
College Enrollment					
UB*Both	0.0342*	0.0212	0.0111	0.0374*	0.0342*
	[0.0140]	[0.0330]	[0.0286]	[0.0167]	[0.0140]
UB*First-Gen Only	-0.0396	-0.0426	-0.0222	-0.0448	-0.0396
	[0.0347]	[0.0284]	[0.0310]	[0.0354]	[0.0347]
UB*Low-Inc Only	-0.0493	-0.00444	0.0238	-0.0683	-0.0493
	[0.0435]	[0.0778]	[0.0890]	[0.0503]	[0.0435]
High School GPA					
UB*Both	0.0308	-0.0141	-0.0283	0.0281	0.0308
	[0.0378]	[0.0930]	[0.0835]	[0.0381]	[0.0378]
UB*First-Gen Only	0.022	-0.06	-0.0193	-0.0034	0.022
	[0.0746]	[0.0946]	[0.0551]	[0.0707]	[0.0746]
UB*Low-Inc Only	0.345***	0.424**	0.411**	0.355***	0.345***
	[0.0881]	[0.143]	[0.133]	[0.0901]	[0.0881]

High School Graduation					
UB Assignment	0.0390*	0.0232	0.0195	0.0409*	0.0390*
	[0.0154]	[0.0398]	[0.0364]	[0.0160]	[0.0154]
UB*First-Gen Only	0.00176	0.0245	0.0327	0.00575	0.00176
	[0.0380]	[0.0411]	[0.0334]	[0.0367]	[0.0380]
UB*Low-Inc Only	-0.00799	0.0461	0.0472	0.00231	-0.00799
	[0.0346]	[0.0560]	[0.0681]	[0.0409]	[0.0346]
Weighting Options	None	SW Full	SW Full	Trim AM	Trim MSE
Site-Spec. Intercepts			Y		

Notes: Specifications are identical to Table 3 but omitting model 1. Covariates included in all specifications. * p < .05, ** p < .01, *** p < .001

Table 9: Predictors of Site Effects

Program Characteristics	*BA+*	*AA+*	*Cert+*	*Enroll*	*HS Grad*
Suburban	-0.2355***	-0.3192***	-0.1960*	-0.1542***	-0.0853+
	(0.0555)	(0.0465)	(0.0843)	(0.0464)	(0.0517)
Urban	0.0199	-0.0633*	-0.0423	-0.0640*	-0.0155
	(0.0297)	(0.0314)	(0.0408)	(0.0285)	(0.0259)
Four-Year Private	0.0996**	0.0826*	0.0654	-0.0287	-0.0115
	(0.0319)	(0.0367)	(0.0407)	(0.0334)	(0.0271)
Two-Year	-0.0036	0.0252	0.0118	-0.0812*	0.0105
	(0.0318)	(0.0411)	(0.0505)	(0.0408)	(0.0418)
Years in Operation (Site)	0.0055*	0.0041+	0.0046	-0.0031	0.0024
	(0.0025)	(0.0022)	(0.0029)	(0.0026)	(0.0023)
FTE Staff per Participant	-0.8349+	-0.8389	-1.1291	0.2101	-0.2471
	(0.4637)	(0.6015)	(0.7337)	(0.3683)	(0.4625)
Avg Years Experience (Staff)	-0.0033	0.0004	0.0028	0.0089	-0.0025
	(0.0050)	(0.0045)	(0.0059)	(0.0054)	(0.0045)
% Typically Ineligible	0.1257	0.1894	0.1845	0.2311	0.1745
	(0.1232)	(0.1629)	(0.1830)	(0.1837)	(0.1572)
# Participants	-0.0027**	-0.0027**	-0.0028*	-0.0001	-0.0012
	(0.0008)	(0.0009)	(0.0011)	(0.0007)	(0.0009)
Performance Requirements	-0.0259	-0.0661	-0.2744***	-0.0294	-0.0133
	(0.0348)	(0.0457)	(0.0501)	(0.0375)	(0.0384)
Types of Services Offered					
# College Prep Services	0.0149	0.0003	-0.0063	0.0467***	0.0117
	(0.0132)	(0.0163)	(0.0194)	(0.0138)	(0.0147)
# Work Prep Services	-0.0298+	-0.0211	-0.0223	-0.0125	-0.0300*
	(0.0181)	(0.0176)	(0.0206)	(0.0140)	(0.0133)
# Self-Awareness Services	0.0385	0.0329	0.0364	-0.0547*	0.01
	(0.0255)	(0.0333)	(0.0383)	(0.0240)	(0.0260)
# Field Trip Types	0.0547+	0.0304	0.0143	-0.0067	0.0233
	(0.0286)	(0.0300)	(0.0377)	(0.0206)	(0.0249)
# Counseling Services	0.0177	0.0661*	0.0406	0.0059	-0.0341
	(0.0252)	(0.0265)	(0.0331)	(0.0210)	(0.0255)

Notes: Regressions based on Stata "mixed" command with linear link and trimmed sample weights. Treatment effects are estimated for each site as random effects and the reported coefficients reflect how the site program characteristics and context variables predict the site effects. The omitted sector is four-year public institutions. "Years in operation" reflects the years the site has operated a UB program. "Average staff experience" is average years of experience weighted by FTE. "% Ineligible" is based on the site-level version of this variable described in the text and Table A2. "Performance requirements" indicates whether sites have performance requirements for students for continued participation. The "Number of services by type" reflects the count of specific services offered under each of the listed categories. Each covariate is included separately also (without treatment interaction) though these coefficients are not shown.

Table 10: Cost-Benefit Analysis

Assumptions	*All Students in Experiment*		*Typically Eligible*		*Typically Ineligible*	
	MPR Table 4 (A) Unadj wgt	*HNM Table 3 (6) Trim wgt*	*HNM Table 6A (3) Unadj wgt*	*HNM Table 6A (6) Trim wgt*	*HNM Table 6B (3) Unadj wgt*	*HNM Table 6B (6) Trim wgt*
Baseline	$9,048	$45,076	$26,661	$46,861	$104,504	$93,303
Use highest returns to degree	$7,886	$50,959	$33,300	$54,363	$108,740	$99,901
(a) Insignif. UB effects to 0	-$7,301	$67,031	$43,607	$53,671	$101,307	$86,896
(b) Lowest returns to degree	$10,363	$39,361	$20,249	$39,521	$100,071	$86,211
(a) and (b) above	-$7,301	$58,233	$33,095	$43,799	$99,373	$85,085

Notes: All estimates of benefits are based on the TOT equivalents of the estimates listed in the tables (and column numbers) at the top of the table. Several other assumptions are common to all the estimates: discount rate (0.035), productivity growth (0.015). In row (a), coefficients are set to zero if $p>0.05$. For the effects by eligibility, we considered the estimates to be significant if they were significant in any one of the three methods used to identify eligible students (see Tables 6A and 6B). See text for discussion about the returns to various credentials.