Attachment B: Documentation of Key Standards Violations in the Mathematica Reports from the National Evaluation of Upward Bound

Submitted in Support of the Request for Rescinding of the WWC Rating of the Mathematica Reports as "Meets Evidence Standards, Without Reservations"

by Margaret Cahalan and David Goodwin

To document the summary material presented in our letter, below we present a brief background concerning the unusual context of this issue followed by a detailed description of the major research standards violations manifest in the Mathematica reports. Attachment C provides additional documentation.

How Did PPSS Staff First Become Concerned--Context and Background?

PPSS concerns first began in 2005 when the Mathematica lead analyst for the fourth follow-up of the long running study sent PPSS tabulations that showed that the "no impact" results were sensitive to only one of the 67 sampled projects.[1] This staff person revealed for the first time to ED that this one project (known as project 69) was carrying fully 26 percent of the total student weights and this meant that students from project 69 had weights that were 40 times those of the lowest project weighted sample members. The staff person was concerned that this project for unknown reasons had large significant negative impacts on college outcomes and given the sample members large weights the results and conclusions of "no impact" for Upward Bound were being driven by this project. Without this project UB clearly demonstrated statistically significant and educationally meaningful positive impacts in the analyses. PPSS staff asked the Mathematica UB project staff to investigate the issue further both in terms of the substantive reasons for the large negative impacts in project 69 and in terms of the unequal weighting issues. PPSS wished to understand what project 69 was doing wrong such that the project was seemingly decreasing rather than fostering college attendance and degree attainment. Given that this project was supposedly representative of the largest number of 4-year BA and above degree granting UB projects—these results could not be ignored. PPSS also wished to understand the role that the uneven weighting issue was playing in the estimates, having observed that the sub-group estimates were unstable and sensitive to issues of differential survey non-response and coverage issues. A few months later, without having gotten a response from Mathematica concerning these issues, the lead analyst working on the fourth follow-up report, left Mathematica. The fourth follow-up report was never revised to address PPSS internal reviewers concerns or put into review for publication[2] .

[1] Details on memo's and emails concerning PPSS concerns can be found in the COE Request for Correction, submitted to ED in 2012 and available on the COE website *The Council for Opportunity in Education (COE) Request for Correction* is available at http://www.coenet.us/files/pubs_reports-COE_Request_for_Correction_011712.pdf

[2] The PPSS review of the fourth follow-up draft raised a number of issues some of which were similar to those of concern raised to the fifth follow up report. However, Mathematica reported they did not have resources to revise the fourth follow up report and also prepare the fifth follow up report. PPSS agreed to Mathematica's concentration on completing the fifth follow up report with remaining contract funds.

The PPSS Quality Assurance (QA) Review

Given the concerns raised, in 2006, ED-PPSS began an internal QA review of the study. The QA review PPSS conducted involved an internal review and analysis of all data files from the study, as well as consultation and replication of PPSS re-analysis results by external statistical experts. Over the period of 2006-2007 as they became available, PPSS obtained from Mathematica copies of the data files through fifth follow-up and had them matched with the federal aid files. Data files reviewed included: the initial sampling frame, the baseline survey, 5 follow-up surveys, student transcripts, 10 years of federal aid files and National Student Clearinghouse (NSC) data.

PPSS also began to seek external statistical advice including consultation with Dr. James Chromy of RTI, the PPSS Technical and Statistical Assistance contractor at the time. Dr. Chromy is a well-respected Fellow of the American Statistical Society who has consulted on sampling issues for NAEP, NPSAS, BPS, B&B in the education field and also has expertise in experimental design issues in health clinical trials. Dr. Chromy and his statistical team gave PPSS advice concerning the sample design issues and also after obtaining the data files from the study, replicated the significant positive impact estimates that PPSS had found internally for the entire sample including Project 69, following the merge with the federal aid files using data files through the fourth follow-up. Unfortunately PPSS was not able to obtain a copy of the grantee population sampling frame from which the UB grantee sample was drawn from Mathematica until December of 2007, after the contract had ended and after the final report had been through an initial review.[3] After PPSS obtained the sampling frame and gained knowledge of the identity of sampled UB projects, PPSS became more concerned about the representational issues in addition to the unequal weighting and treatment and control group lack of balance.[4] The fact that the Project 69 was a former minority serving junior college and had an historical focus on certificates and service to CTE high schools was also helpful in understanding the large impacts on certificate attainment that had been observed. Gradually, over the period of 2006 to 2009 as study errors were identified, PPSS staff used NCES and WWC standards to mitigate these errors in a re-analysis described in more detail below.

Major Standards Violations and Their Impact

As already noted in our letter, listed below are the key applicable sets of standards and guidelines used in identifying key flaws in the Mathematica Upward Bound reports.

- *U.S. Department of Education Information Quality Guidelines (ED Guidelines)*
- *Joint Committee on Standards for Educational Evaluation (JCSEE).* http://www.jcsee.org/)

[3] At the request of Dr. Chromy of RTI, the PPSS Technical and Analysis Support Contractor at the time, whom PPSS had asked to provide an independent review of the Mathematica UB sample design, PPSS requested the population sampling frame from Mathematica in March of 2007. However, due to delays in Mathematica's ability to locate the population frame it was 10 months before it was finally delivered to PPSS and after the contract had ended in December of 2007. Prior to the end of the contract in December of 2007, PPSS did not have knowledge of the identity of project 69.

[4] More recently, Dr. Chromy also reviewed the COE Request for Correction (2012) and gave advice concerning recommendations for mitigation. *The Council for Opportunity in Education (COE) Request for Correction submitted in 2012* is available at http://www.coenet.us/files/pubs_reports-COE_Request_for_Correction_011712.pdf,

- *IES, National Center for Education Statistics (NCES) Statistical Standards*--- These may be accessed at the following site url: http://nces.ed.gov/statprog/
- *IES, What Works Clearinghouse Standards (WWC)* ---these may be accessed at the following site urls: http://ies.ed.gov/ncee/wwc/pdf/wwc_version1_standards.pdf
- http://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1/
- *American Educational Research Association* (AERA) *Standards for Reporting on Empirical Social Science Research* in AERA Publications http://www.sagepub.com/upm-data/13127_Standards_from_AERA.pdf

Exhibits B-1 and B-2, (repeated from our letter) respectively first identify key specific applicable standards that have been violated and then summarize the 10 specific violations in the Mathematica reports. There follows a detailed discussion of each of the 10 violations with supporting exhibits documenting errors and re-analyses findings. Attachment C under a separate file in this package contains additional documentation and includes actual output from STATA tabulations taken from the 2009 report *Addressing Study Error in the Random Assignment National Evaluation of Upward Bound: Do the Conclusions Change? By Margaret Cahalan* a COE report published in 2009 and available at http://www.pellinstitute.org/publications-Do_the_Conclusions_Change_2009.shtml

Exhibit B-1
Key Information Quality Guidelines and Standards that are Applicable to the Concerns with Regard to the *Mathematica Upward Bound Report* s

Department of Education Quality Information Guidelines

Research and Evaluation information products should, at a minimum: …

- Pose the research or evaluation question in a balanced and unbiased manner;
- Provide an unbiased test of the question; …
- Present conclusions that are strongly supported by the data; ….
- Confirm and document the reliability of the data, and acknowledge any shortcomings or explicit errors in any data that is included;
- The source of data should be reliable. The sample should be drawn from a complete list of items to be tested or evaluated, and the appropriate respondents should be identified, *correctly sampled,* and queried with survey instruments that have been properly developed and tested
- Appropriate steps should be taken to *ensure that the respondents are a representative sample*;

What Works Clearinghouse Handbook of Procedures and Standards

A study may fail to meet WWC evidence standards if ..…….

- It does not include a valid or reliable outcome measure, or does not provide adequate information to determine whether it uses an outcome that is valid or reliable. …..
- The intervention and comparison groups are not shown to be equivalent at baseline
- The overall attrition and or differential attrition rate exceeds WWC standards for an area.
- The measures of effect cannot be attributed solely to the intervention

……

NCES Statistical Standards Concerning Non-Response and Coverage

- **STANDARD 2-2-4:** A nonresponse bias analysis is *required* at any stage of a data collection with a unit response rate less than 85 percent. The extent of the analysis must reflect the magnitude of the nonresponse (see Standard 4-4).
- **STANDARD 3-1-2:** NCES data collections that are used as sampling frames for other NCES surveys must strive for coverage rates in excess of 95 percent overall and for each major stratum. **STANDARD 3-1-3:** If there is not evidence of a coverage rate of at least 85 percent of the target population, then frame enhancements such as frame supplementation or dual frame estimation must be incorporated into the survey study design.

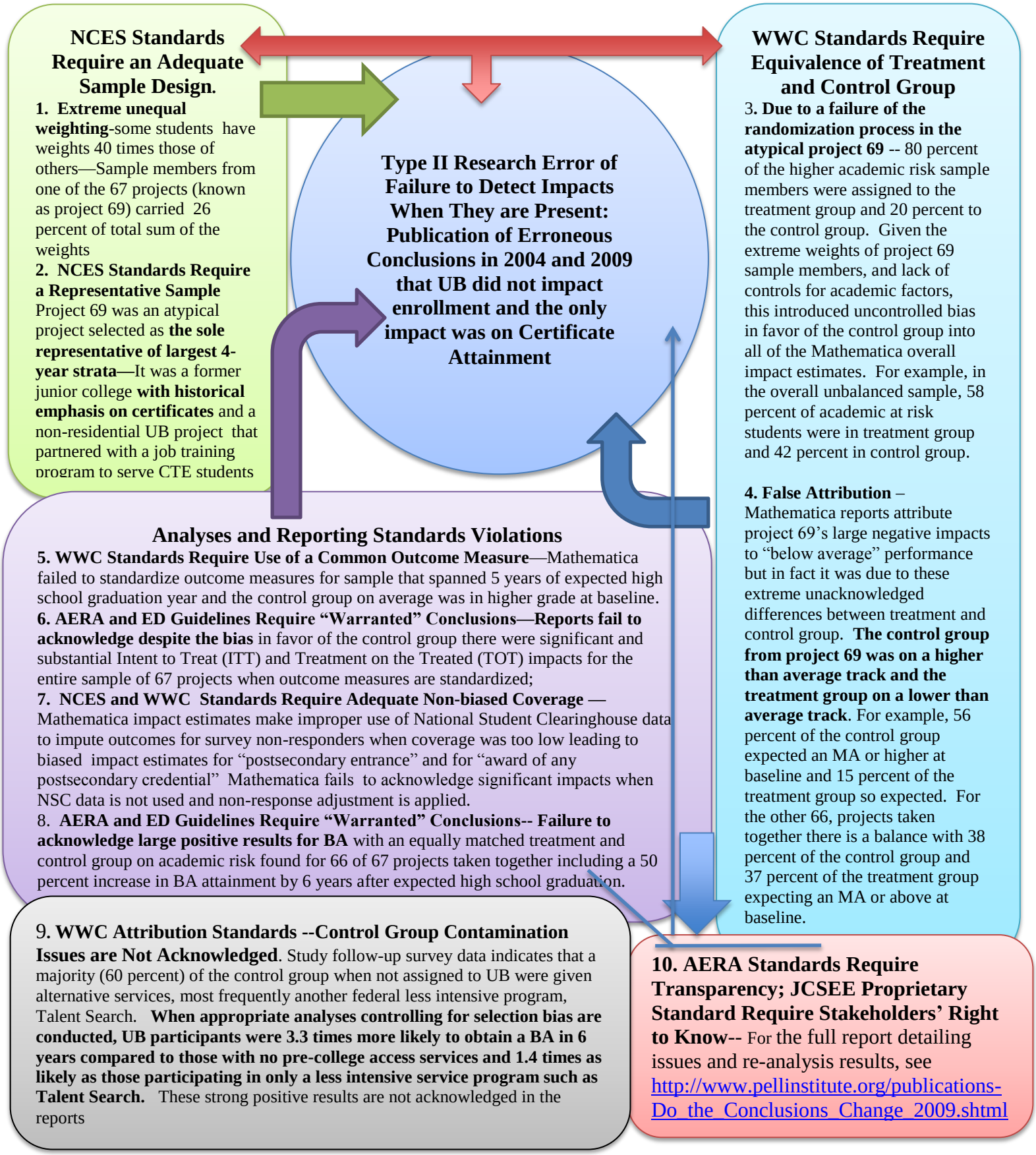
Joint Committee on Standards for Education Evaluation Standards: The Joint Committee Standards address ethics of research under the heading of Propriety. Standard P6 noted below discusses the full disclosure of findings

- **P6 Disclosure of Findings** The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation and any others with expressed legal rights to receive the results

American Educational Research Association (AERA) Standards for Reporting on Empirical Social Science Research

- Two overarching principles underlie the development of these reporting standards: the "sufficiency of the warrants" and the "transparency" of the report.

Exhibit B-2. NCES, WWC, JCSEE, AERA Standards and ED Guidelines, Violations in the Mathematica Upward Bound (UB) Evaluation Reports



Violation 1: **Using Flawed Sample Design to Make Inferences of Average Impact**. *NCES Statistical Standards and Department of Education Information Quality Guidelines require that the sampling follow correct procedures, that it be checked and found to be sufficient for robust estimation of the population for which the study is intended to generalize; (NCES Statistical Standards; commonly accepted procedures for inverse probability of selection weighting procedures).*

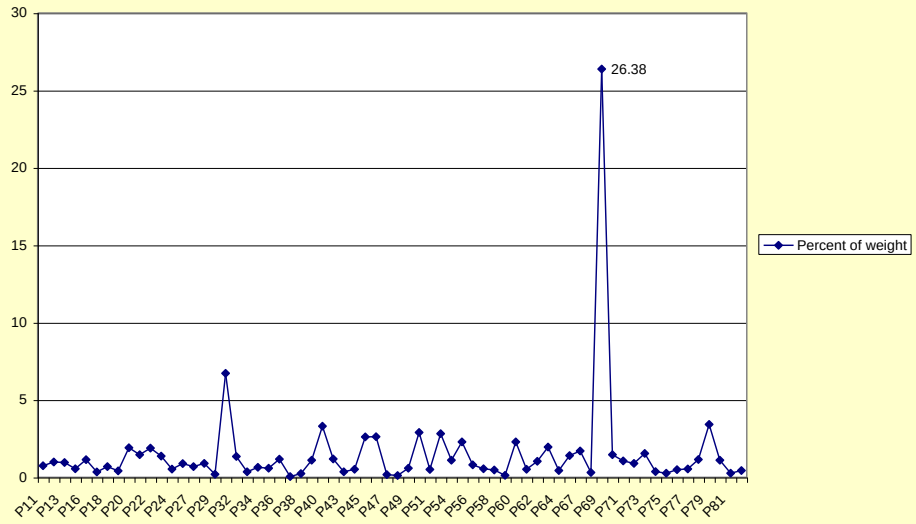
In a sample design that internal and external reviewers have consistently called "seriously flawed" and which violates NCES standards, the multi-stage sample design had only one single project selected to represent the largest study defined 4-year and above stratum. As Dr. Chromy has noted, "If representing the stratum of four-year institutions was a clearly stated objective of the study, than a sample of size 1 for this group is clearly inadequate." In the final stage of weighting this one project of 67 ended up carrying fully 26 percent of total student sample weights (Exhibit B-3). As noted, this meant that students sampled from this UB project had weights 40 times those of students in the lowest weighted UB projects.

An IES blind external reviewer summarized the issue with the following language asserting that inferential estimates could not be considered robust with such a design.

> *"The decision made in 1992-94 to select only a single project at random from this 26% share of the applicant population created a design in which design-based estimation and inferences for the full population could not be robust for the true population values. Simply applying a population weight to an inadequate sample of one cluster from a 26% share stratum will not correct this.*
>
> *What to do? With respect to design-based inferences for all other strata, the baseline sample of programs should enable robust inferences for that share of the UB survey population not included in the Project 69 stratum." (IES external statistical reviewer C July 2008*)

All of the estimates in the 2004 and 2009 Mathematica reports were based on non-robust impact estimates that included Project 69. Impact estimates representing 74 percent of the intended population that had been identified as potentially more robust were not considered by Mathematica in the determination of conclusions they published. During preparation of the fifth and final UB report, PPSS Technical Monitors repeatedly recommended that Mathematica report UB impacts with and without project 69 and with and without the study weights and that the error issues with project 69 be considered in drawing conclusions about Upward Bound. These recommendations were repeatedly refused by Mathematica. The 2004 report, written before ED-PPSS knew of or raised questions concerning the limitations of the sample design, does not even mention the unequal weighting issue at all. The 2009 report does so only in a non-transparent manner that does not acknowledge the seriousness of the issue, or the equally significant if not more significant representational and lack of treatment and control group balance issues associated with project 69 discussed below in Violations 2 and 3.

Exhibit B-3. Percentage distribution of sum of the weights by project for the 67 projects making up the study sample: National Evaluation of Upward Bound, study conducted 1992-93-2003-04



SOURCE: Data tabulated April 2009 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education; study conducted 1992-93 to 2003-04.

Violation 2. NCES Standards and ED Information Quality Guidelines Require a Representative Sample. *The Mathematica reports use an "atypical project" as sole representative of largest public 4-year and above stratum. NCES standards specify that the sample be checked to see that it is representative of the population of interest*.

As PPSS career staff found out after the contract had ended, Mathematica had randomly selected "an atypical for its stratum" Upward Bound project to be the sole representative of the largest public stratum hosted at 4-year BA and above granting institutions. Dr. Cahalan and Dr. Goodwin together found when Mathematica finally sent the sampling frame list of eligible grantees to ED and the identity of project 69 became known to them at the end of the contract in December of 2007, that the project with 26 percent of the weight that was supposedly the sole representative of the largest 4-year stratum, although officially classified as a 4-year college was in fact a former minority serving junior college that had been taken over by a city college system to serve as their downtown campus. The UB program partnered with a job-training program and the grantee institution had historically had a large number of certificate programs linked to the CTE high schools served. It also did not have the hallmark 4-year grantees' UB summer residential program present in virtually all projects within the 4-year stratum it was representing—as it has no on-campus housing. The large diverse frame stratum for which project 69 is the sole representative included major flagship research universities that had UB grants at the time as well as public small 4-year majority white and majority black liberal arts colleges. None of these types of UB programs could be adequately represented by project 69.

This representation issue combined with the large weight and the serious lack of balance between the treatment and control group (to be discussed below) resulted in a probable over-estimate of the UB impact on certificate attainment and importantly a serious underestimate of the impact on BA receipt. In the context of recommending what should have been done at the start of the study external reviewer Dr. Chromy states:

> "With an imperfect sampling frame, it would be an accepted practice to check each project drawn and drop it from the sample if it does not meet the study population definition; this is a form of screening for eligibility. (Dr. James Chromy comments on the COE Request for Correction, October 2011)."

After the contract had ended, but while the final report was still under review, both Dr. Cahalan and Dr. Goodwin in separate communications with Mathematica asked Mathematia to acknowledge these concerns in the revisions to the report. Mathematica refused to do this and in fact sentences were inserted in the Executive Summary in the final changes negotiated with the political appointees that misled readers into thinking that project 69 was an adequate representative of the 4-year and above stratum. In the 2009 report Executive Summary Mathematica specifically states:

> *"Project 69 was similar to other projects in this stratum on a broad range of characteristics. Similarly data from the student surveys and NSC and FSA records indicated that the students from project 69 did not have unusual characteristics" (Executive Summary Mathematica Fifth Follow-up Report, Page xvii- xviii**).***

The Mathematica 2009 report goes on to state that analyses that omit project 69:

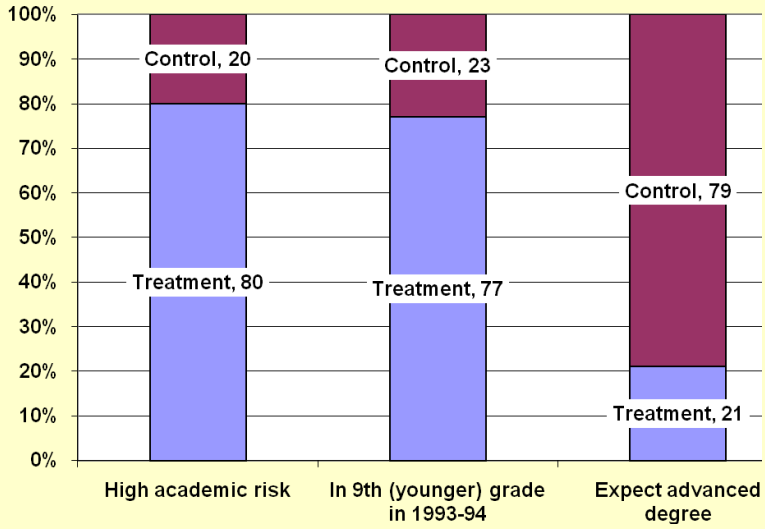
> *Do not appropriately represent the most common stratum of Upward Bound projects. Thus such analyses do not answer the evaluation's research questions about the impacts of the national Upward Bound program. Moreover the estimates for such analyses do not generalize to urban projects, large projects or any other well-defined subset of projects for which the findings might have policy implications. In contrast the findings from the main impact analyses, which include all projects weighted based on their selection probabilities are intended to generalize to the national Upward Bound program" (Executive Summary, Mathematica Fifth Follow-up Report, page xviii)*

Violation 3: WWC Standards Require Equivalence of Treatment and Control Group. *The Mathematica reports violate basic WWC and common random assignment study standards by using a seriously unbalanced non-equivalent treatment and control group to estimate Upward Bound impact and by not acknowledging the importance of this lack of balance in the report.*

WWC Standards require that the treatment and control group are equivalent on factors related to outcomes at the baseline and throughout the study. In what some external reviewers identified as a

probable failure to implement the random assignment correctly, PPSS found that the atypical and highly weighted project 69 also had a severe lack of balance between the treatment and control group. This lack of balance combined with the extreme weights resulted in an uncontrolled bias in favor of the control group in all of the Mathematica impact estimates as well as a lack of stability of estimates subject to non-response and non-coverage issues. As noted in project 69, the control group on average was found to be much more academically proficient, to be in a higher grade at baseline, and had much higher educational expectations than the treatment group (Exhibit B-4). Mathematica had no controls for academic risk indicators in their models and inadequate controls for grade at a fixed time and educational expectations. In addition as discussed under Violation 5, the utility of these later controls was compromised by the fact that outcome measures used were not standardized to expected high school graduation date. Due to the fact that students can enter Upward Bound from rising 9th graders to rising 12th graders, and the fact that recruitment for the study spanned 2 calendar year summers, the sample spanned 5 years of expected high school graduation year cohorts.

Exhibit B-4. Percentage distributions for project 69 between treatment and control groups among those sample members who were a higher academic risk, in 9th (earlier) grade in 1993-94, and who expected an advanced degree at baseline: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04 (N= 85)



Note large imbalance in project 69 distribution. Figure is read as follows: For example, among those who were classified as higher academic risk, 80 percent were in the treatment group and 20 percent in the control group. In a random assignment study distribution should be 50-50 between treatment and control group; figure shows imbalance in project 69.

NOTE: High academic risk includes those sample members in the bottom 20 percent of the sample on 9th grade GPA and other academic indictors. . There were not controls in the models for academic variables. **SOURCE:** Data tabulated April 2009 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education; study conducted 1992-93 to 2003-04.

The PPSS balance checks found that the treatment group sample resembled on average the students from an historically serving minority Career and Technical Education (CTE) Target high school served by the certificate granting project 69 institution grantee. On the other hand the control group from project 69 on average resembled the applicants to Upward Bound Math Science (UBMS) Initiative —known to be beginning at the time in the area under a separate UBMS grant (not selected to be in the study sample). PPSS suspected that the unusually large number of "applications" (baseline surveys submitted by this project) included baseline surveys for those more academically proficient and higher grade students interested in the newly initiated UBMS project. It is probable that these were not in fact "applicants" to the project 69 CTE focused Upward Bound program and should have been excluded from the study "waiting list sample" as ineligible.

Exhibit B-5 further illustrates the differences between the treatment and control group in project 69 with the treatment group resembling CTE students and the control group resembling on average students similar to Upward Bound Math Science applicants. For example, among the 66 other projects taken together about 38 percent of the control group and 37 percent of the treatment group reported at baseline that they expected an MA degree or higher. In contrast in project 69, among the control group 56 percent expected an MA degree or higher and among the treatment group 15 percent expected an MA degree or higher at baseline.

In project 69 the control group which consisted of all those who had completed a baseline survey and who were not randomly selected to be invited into the program as openings occurred was substantially larger than the treatment group. This was because the number of openings was less half the unusually large number of baseline surveys submitted by project 69. It should also be noted that within project 69 there were two project defined sub-strata so there was unequal weighting within project 69 itself, as most of the random assignment openings were filled from one of the two strata. This factor also contributed to the estimation issues from this project's sample. Whatever the reason they occurred, the substantial differences between the treatment and control group in project 69 accounted for the large negative impacts found in this project when it was considered alone.

Exhibit B-5. Percentage of project 69 and all other projects having various attributes by treatment and control group status: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04

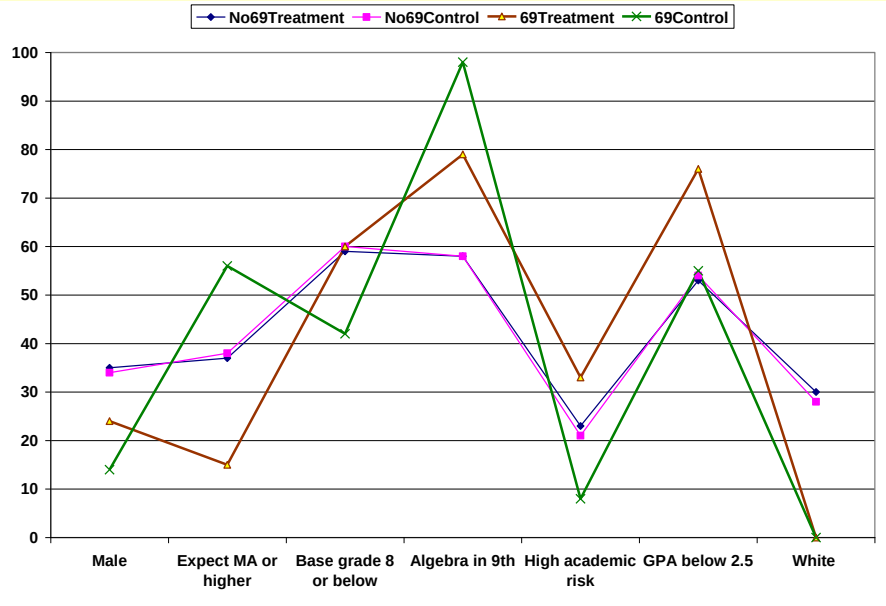


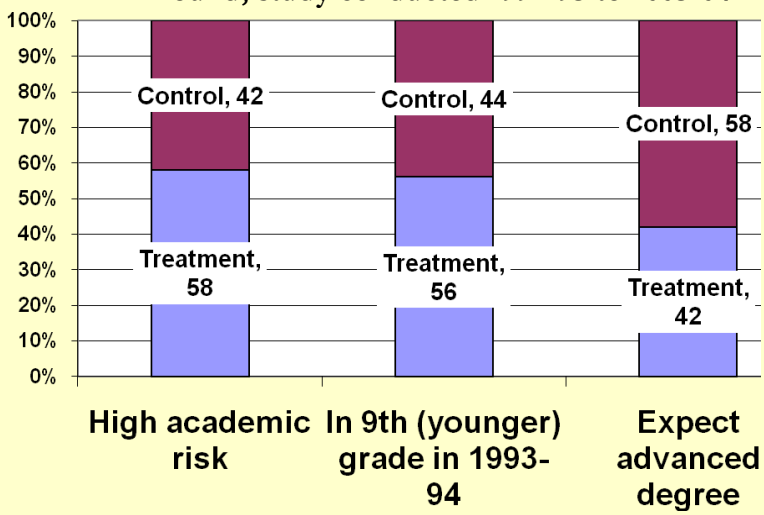
Figure shows that the UB treatment and control group are well matched without Project 69 on the variables in the chart; however, in project 69 the treatment and control group manifest substantial differences. For example, 56 percent of the control group in project 69 expected an MA or higher at baseline compared with 15 percent of the treatment group. In contrast, among the other 66 projects in the sample, 38 percent of the control group and 37 percent of the treatment group expected an MA or higher.

NOTE: Project 69 tabulation based on the 85 sample cases from project 69 (52 controls and 33 treatment cases -- poststratified weighted to 11,536 cases -- 5,768 treatment and 5,768 controls). The category "No69treatment" and "No69control" represents all the other projects in the sample excluding project 69; these 66 projects are considered to represent 74 percent of the UB applicants in the study period.

SOURCE: Data tabulated December 2007 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education; study conducted 1992-93 to 2003-04.

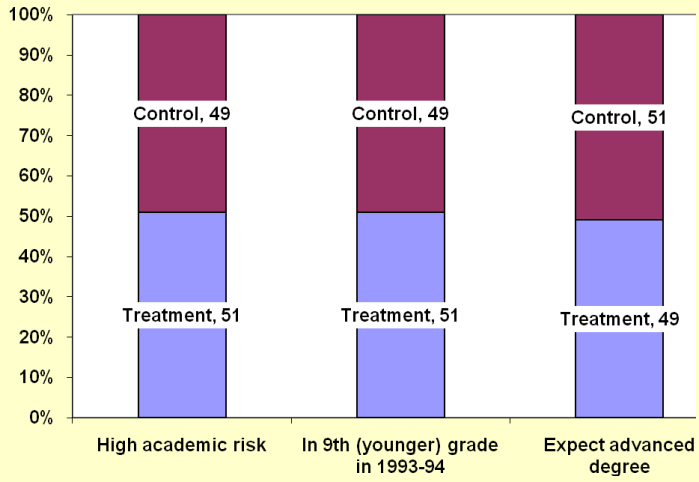
As noted, the academic risk differences between the treatment and control group were uncontrolled for in the analyses and introduced serious bias in favor of the control group into all of the Mathematica overall impact estimates (Exhibit B-6). As can be seen in Exhibit B-7, the other 66 projects when taken together exhibit a reasonable balance on these factors between the treatment and control group as one would expect in a random assignment study. For these reasons external reviewers such as Dr. Chromy and the IES blind external reviewer noted above recommended that the study acknowledge these limitations and attempt to make inferences only based on 66 of the 67 projects in the sample that exhibited a reasonable balance between treatment and control and a reasonable representation—especially in considering the impact on BA attainment.

Exhibit B-6. Percentage distributions in all 67 sampled projects (including project 69) between treatment and control groups on various attributes: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04



Note with project 69 included **there is not a balance or equivalence between treatment and control group** as WWC standards require in a random assignment study. Figure is read as follows: For example, among those who were classified as higher academic risk, 58 percent were in the treatment group and 42 percent in the control group. In a random assignment study distribution should be about 50-50 between treatment and control group; figure shows imbalance in overall sample with project 69 included.

Exhibit B-7. Percentage distributions for 66 of 67 sampled projects (excluding project 69) between treatment and control groups among those sample members who were a higher academic risk, in 9th (earlier) grade in 1993-94, and who expected an advanced degree at baseline: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04



Note without project 69 there is a balance between treatment and control group as expected in random assignment study. Figure is read as follows: For example, among those who were classified as higher academic risk, 51 percent were in the treatment group and 49 percent in the control group. In a random assignment study distribution should be about 50-50 between treatment and control group.

NOTE: High academic risk includes those sample members in the bottom 20 percent of the sample on 9th grade GPA and other academic indictors. . There were not controls in the models for academic variables

SOURCE: Data tabulated April 2009 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education; study conducted 1992-93 to 2003-04.

Violation 4: False Attribution of Project 69's negative impacts when considered alone ---*What Works Clearinghouse Standards require that the impacts be attributable to the intervention being studied***.**

Mathematica's failure to detect impacts and type II errors are related to a number of factors, but as seen in Exhibits B-4 to B-7, one factor in their error was attributing a lack of impact findings as related to the Upward Bound program performance and not to the uncontrolled bias present due to the lack of balance between the treatment and control group on academic factors. The Mathematica reports specifically has misleading attribution of "below average impacts" for project 69, in implying that project 69 had below average impacts due to the Upward Bound program's "below average" (poor) performance when in fact these negative impacts were due to the gross lack of equivalency between the treatment and control group in this project. The Mathematica report states: "*Because Project 69 had below average impacts, reducing its weight relative to other projects resulted in larger overall impacts for most outcomes compared with the findings from the main impact analysis, which weighted all sample members according to their actual selection probabilities."* This is a misleading statement about the effectiveness of project 69. As noted above in Exhibit B-4, a closer look at project 69's treatment and control group clearly reveals that the so-called "below average" (negative impacts) in this project were not due to project 69's "poor performance" but were due in fact to the extreme differences between the treatment and control group in favor of the control group in this project.

As noted above, the facts concerning project 69's representational issues and lack of treatment and control group balance issues are not acknowledged in the Mathematica reports. This misrepresentation of a largely 2-year and less than 2-year grantee as the sole representative of the largest 4-year public stratum, combined with the extreme large weight and the uncontrolled for academic factor bias in favor of the control group in this project's sample contributes to a probable type I error of over-estimating the impact of UB on the attainment of certificates of the type awarded by the project 69 grantee institution and a type II error of failure to detect effects for the attainment of Bachelor's degrees. Very strong positive effects on BA attainment were found for 66 of the 67 sampled projects taken together, which were found to meet WWC standards for baseline equivalence and were well matched when taken together on relevant attributes (see discussion below on BA attainment), but not when project 69 is included in the impact estimates.

Violation 5. Failure to Standardize Outcome Measures. *WWC and NCES standards require that any comparisons use standardized and precise common outcome measure. Mathematica did not do this in any of its outcome measures used to assess UB impact.*

As noted, the UB student sample spanned 5 years of expected high school graduation years. When PPSS reviewers asked that Mathematica standardize the outcome measures, they argued that their random assignment method made standardizing by expected high school graduation year unnecessary. Consequently Mathematica failed to standardize their outcome measures to differences in expected high school graduation years in either the 2004 or 2009 reports. PPSS QA review found the control group on average to be in a higher grade at a fixed time period, and this contributed to the uncontrolled bias in the Mathematica impact estimates for key postsecondary outcome measures (enrollment, financial aid,

postsecondary degrees and credentials). The lack of standardization of outcome measures also decreased the utility of the control variables used in the regression models. After Mathematica did not respond positively to PPSS and OPE's suggestions that the outcome measures be standardized and after the contract had ended, in spring of 2008, PPSS internal staff derived the dates necessary to standardize the primary outcome measures. Exhibit B-8 below shows the differences between unstandardized impact estimates and standardized impact estimates using the third follow-up survey data and administrative data.

Exhibit B-8. Various model results using third follow-up survey responders only and using full longitudinal sample for evidence of entering postsecondary for ITT and TOT models: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04

	All sampling strata		**One project with bias removed (the remainder represents 74 percent of Horizons waiting list)**	
	Given Opportunity (ITT)	**Participated in UB/UBMS (TOT/CACE)**	**Given Opportunity (ITT)**	**Participated in UB/UBMS (TOT/CACE)**
Third follow-up survey responders only with no administrative records and no standardization of outcome to expected high school graduation year (EHSGY); uses non-response adjusted weight	pr-T = 76.4 pr-C = 75.4 Difference = 1.0NS	xb T = 75.4 xb C = 71.7 Difference = 3.7 NS	pr T = 77.8 Pr C = 72.2 Difference = 5.7**	xb T = 77.6 xb C = 67.7 Difference = 9.9*
Third follow-up survey responders only with no administrative records or other applicable surveys, but with standardization to +1 (18 months) of expected high school graduation year; uses non-response adjusted weight	pr-T = 71.2 pr-C = 68.2 Difference = 3.0 NS	xb T = 71.4 xb C = 65.2 Difference = 6.1 NS	pr T = 73.3 Pr C = 65.8 Difference = 7.5***	xb T = 74.0 xb C = 61.9 Difference = 12.1***
Third follow-up survey responders only – standardized to +1 (18months) of EHSGY and uses all applicable surveys and Student Financial Aid (SFA) records; uses non-response adjusted weight	pr-T = 75.9 pr-C = 71.4 Difference = 4.6*	xb T = 76.0 xb C = 67.8 Difference = 8.2 NS.11	pr T = 77.8 Pr C = 70.0 Difference = 7.8****	xb T = 78.2 xb C = 65.6 Difference = 12.6***
Includes all sample members, standardized to +1 (18months) of EHSGY and uses all applicable surveys and SFA records; uses poststratified adjusted weight	pr-T = 72.9 pr-C = 66.0 Difference = 6.9****	xb T = 73.5 xb C = 62.5 Difference = 10.9****	pr T = 73.3 Pr C = 64.3 Difference = 9.1***	xb T = 74.6 xb C = 60.4 Difference = 14.2****

*/**/***/**** Significant at 0.10/0.05/.01/00 level; NS = not significant at the .10 level or below. UB = regular Upward Bound; UBMS = Upward Bound Math/Science; ITT = intent to treat; TOT = treated on treated**;** CACE = complier average causal effect; T = treatment; C = control or comparison; pr = estimated probability from STATA logit regression; xb = linear prediction from STATA ivreg (instrumental variables regression). **SOURCE:** Data tabulated December 2007 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), US Department of Education: study conducted 1992-93 to 2003-04; and federal Student Financial Aid (SFA) files 1994-95 to 2003-04.

The highly influential Third Follow-Up report (Myers, et. al 2004) reporting no impacts on postsecondary outcomes for Upward Bound did not standardized outcomes by expected high school graduation year or use administrative records in addition to the surveys. All impact estimates included the biases introduced by project 69 and no estimates without project 69 were reported. Nor did the report acknowledge that the "no impact" estimates were being driven by one project with large negative impacts. The 2004 report did not reveal any of the bias issues surrounding project 69's extreme unequal weighting, representational issues or the non-equivalent treatment and control group resulting in uncontrolled bias in favor of the control group. (the issues discussed above in violations 1 to 4). The Third Follow-up findings or lack of findings formed the formal justification of the OMB PART rating of "Ineffective" and the Bush Administration's budget requests in FY2005 and FY2006 to zero fund Upward Bound, Upward Bound Math Science, Talent Search and GEAR UP.[5]

Violation 6: Failure to Acknowledge and Report Positive Findings When Standardization Was Implemented. *AERA Standards, ED Information Quality Guidelines, and Joint Committee Standards for Education Evaluation (JCSEE) all require that conclusions be "warranted"*

PPSS review comments to the draft fifth follow-up report in 2007 (Sefter et.al. 2009) had asked for standardization of the outcome measures to expected high school graduation cohort year. However Mathematica continued to argue that this was unnecessary due to their randomization process. As noted in response in early 2008, after the contract ended PPSS internal staff did the standardization and sent a paper to Mathematica in the spring detailing the results that included the output from the STATA runs. Examples of these positive impact estimates are graphed in Exhibits B-9 and B-10 and documentation of the output from the regression models is included Attachment C. Mathematica choose to ignore these positive findings and to engage in a debate about the Baseline Survey variable PPSS used to derive expected high school graduation year. In response PPSS used a different survey variable from the First Follow up Survey and found the same significant and substantial positive impact estimate results which are also included in Attachment C. [6]

The PPSS impact estimates using two different methods to estimate expected high school graduation year both showed significant and substantive positive impacts for Upward Bound with and without project 69 for enrollment and application for and award of financial aid . As noted, despite this information Mathematica chose to not report or even acknowledge that these documented findings had been sent to them in their revisions to the Fifth Follow-up report. Moreover, despite the fact that both Dr. Cahalan and Dr. Goodwin had repeatedly questioned Mathematica's methods and conclusions, Mathematica thanked Dr. Goodwin and Dr. Cahalan in the acknowledgements to the report with no mention of the fact that both Dr. Goodwin and Dr. Cahalan had communicated to Mathematica that they

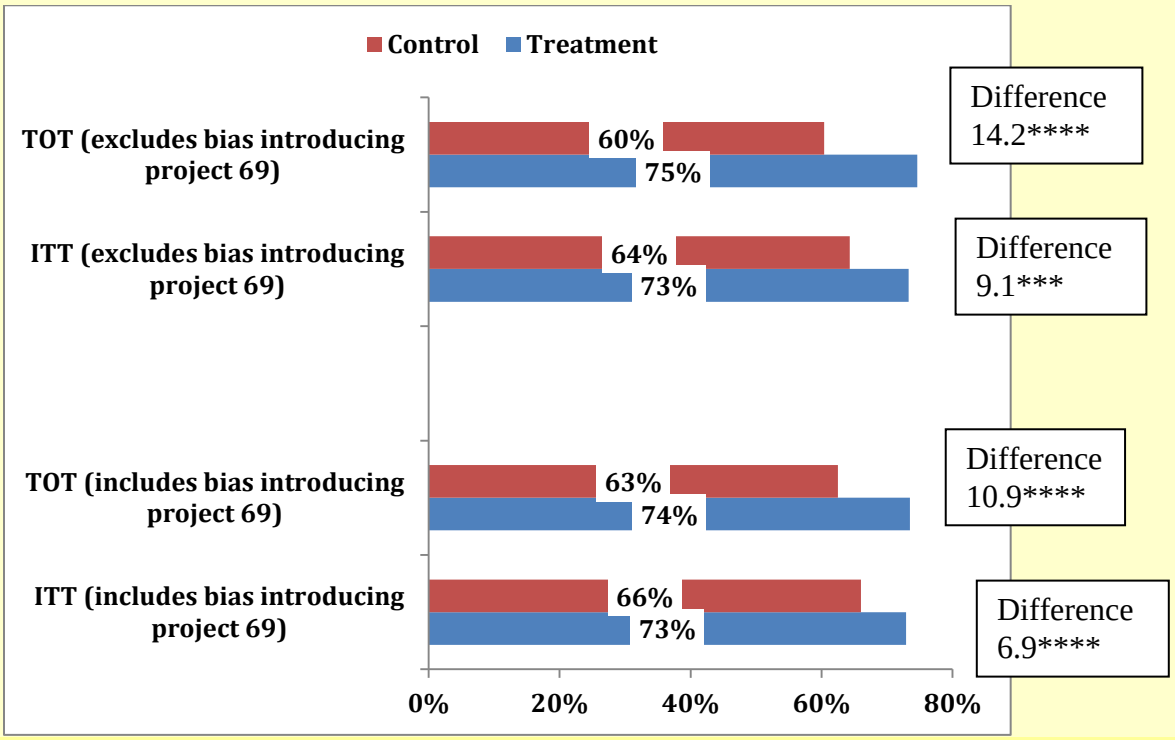
[5] The Third Follow Up results of no overall impact but very large impacts on postsecondary entrance for those students classified as the highest academic risk, also led to the efforts of OMB and the Department of Education to re-focus the Upward Bound program away from preparation for BA attainment and toward serving those students who were considered to be the most at academic risk in their high schools (defined as students who had failed a high school competency test or had very low GPAs) though the Upward Bound Initiative.

[6] Attachment C is taken from appendices from a paper prepared by Cahalan that includes impact estimates from using two different variables to ascertain expected high school graduation year. Cahalan found that the impact estimates were similar when the two different survey questions were used. This exercise was undertaken when Mathematica questioned the variable used by PPSS to ascertain expected graduation year.

viewed the draft report and its conclusions as "seriously flawed." This lack of mention of the positive impacts when results were standardized and lack of mention of the PPSS monitoring staff concerns, gave the misleading impression to readers and reviewers that the concerns with the report conclusions raised by PPSS were unfounded or had been adequately addressed. As can be seen from the examples in Exhibits B-8 to B-10 there is clear documented evidence of positive postsecondary impact when results are standardized by expected high school graduation year for ITT and TOT estimates with and without the bias introducing project 69.

Exhibit B-9 summarizes standardized results for enrollment within +1 year after expected high school graduation year. Similar findings were found for enrollment in +4 year after high school. Exhibit B-10 presents federal aid application within 4 years of expected high school graduation year and is based solely on the administrative records from 10 years of federal aid files (not subject to survey or coverage response bias). Contrary to what is reported in the Mathematica fifth follow up report there is clear evidence of impact on both application and award of aid when aid results from 10 years of federal aid files are standardized by expected high school graduation year.

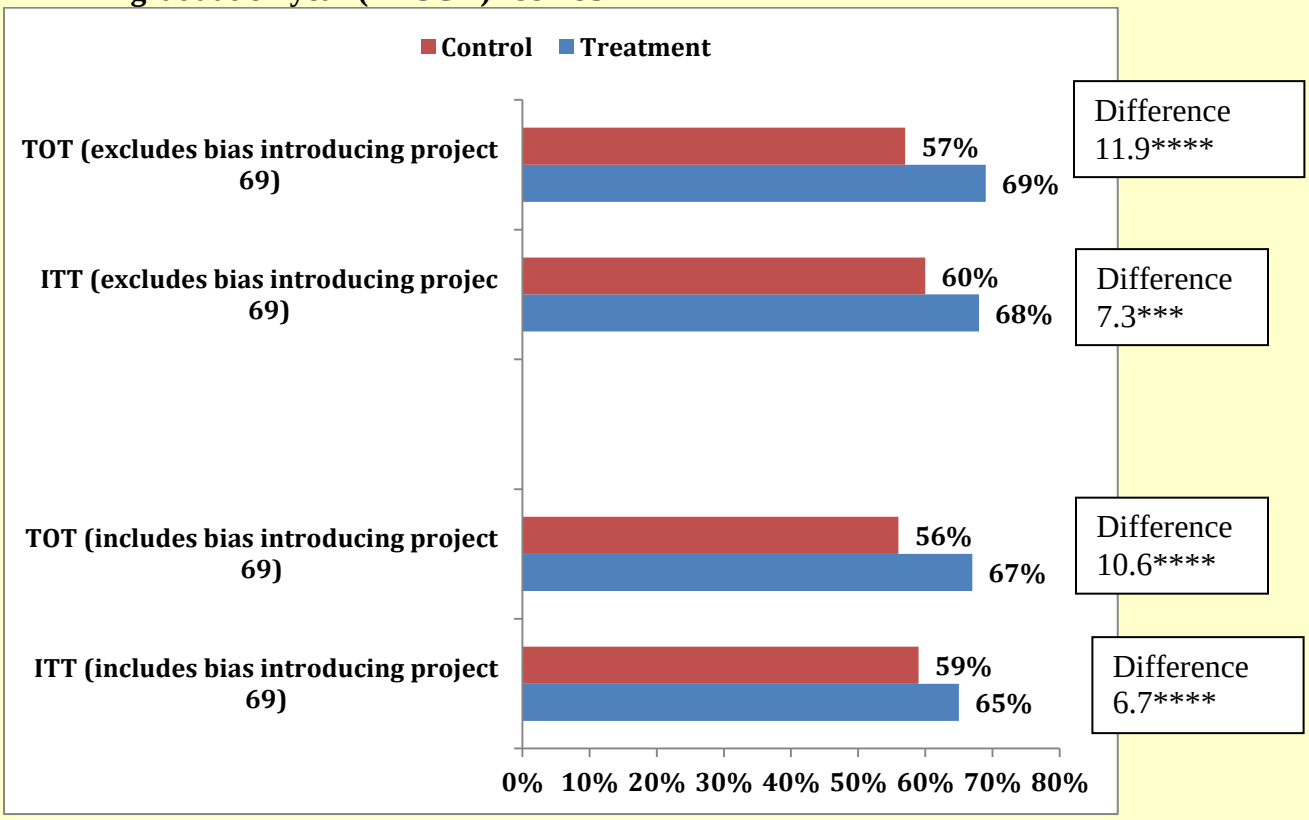
Exhibit B-9. Treated on the Treated (TOT) and Intent to Treat (ITT) estimates of impact of Upward Bound (UB) on postsecondary entrance within +1 year (18 months) of expected high school graduation year (EHSGY) 1992-93 to 2003-04



*/**/***/**** Significant at 0.10/0.05/. 01/00 level.

NOTE. Model based estimates based on STATA logistic and instrumental variables regression and also taking into account the complex sample design. Based on responses to 5 follow-up surveys and federal student aid files. **SOURCE:** Data tabulated January 2008 using: National Evaluation of Upward Bound data files, and federal Student Financial Aid (SFA) files 1994-95 to 2003-04. (Excerpted from the *Cahalan Re-Analysis Report*, Figure IV)

Exhibit B-10. Treated on the Treated (TOT) and Intent to Treat (ITT) estimates of impact of Upward Bound (UB) on federal aid application within +4 years of expected high school graduation year (EHSGY) 1992-93



*/**/***/**** Significant at 0.10/0.05/. 01/00 level;;**NOTE:** Estimated rates from STATA logistic and instrumental variables regression taking into account the complex sample design. Weighted data use poststratified weights. Based on 10 years of financial aid administrative record data. **SOURCE:** Data tabulated January 2008 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education: study conducted 1992-93 to 2003-04; and federal Student Financial Aid (SFA) files 1994-95 to 2003-04

.**Violation 7. Coverage and Bias Issues in the Improper uses of National Student Clearinghouse (NSC) data.** *Both NCES and WWC standards affirm that the study must achieve adequate overall and differential coverage and response, and must concern itself with overall attrition and with differential attrition, non-response and non-coverage bias issues for all of the data sources used relative to the population of interest.*

In the fifth follow-up report Mathematica violated NCES coverage standards in making improper use of the NSC data to impute all fifth follow-up survey non-responders (about 25 percent of the sample) not found on the NSC files to negative values for the outcomes of postsecondary enrollment, and award of any degrees. The applicable time period was when NSC reported only 26 percent coverage for enrollment and when 2-year and below degree data was not yet even being collected by NSC. There was also evidence of bias due to the fact that the heavily weighted project 69 did not participate in NSC

until a date after its sampled target school graduates would have graduated high school. There is evidence that this improper use of the NSC data for the fifth follow-up non-responders combined with the unequal weights increased sensitivity to small variations in coverage and response and led to erroneous conclusions from the study for enrollment estimates but also especially for the conclusions with regard to the important output measure of "award of any postsecondary degree or certificate."

As the statement below taken from the Executive Summary notes, Mathematica reported that the study detected no impact on award of "any postsecondary degree or credential." The 2009 Mathematica report executive summary states:

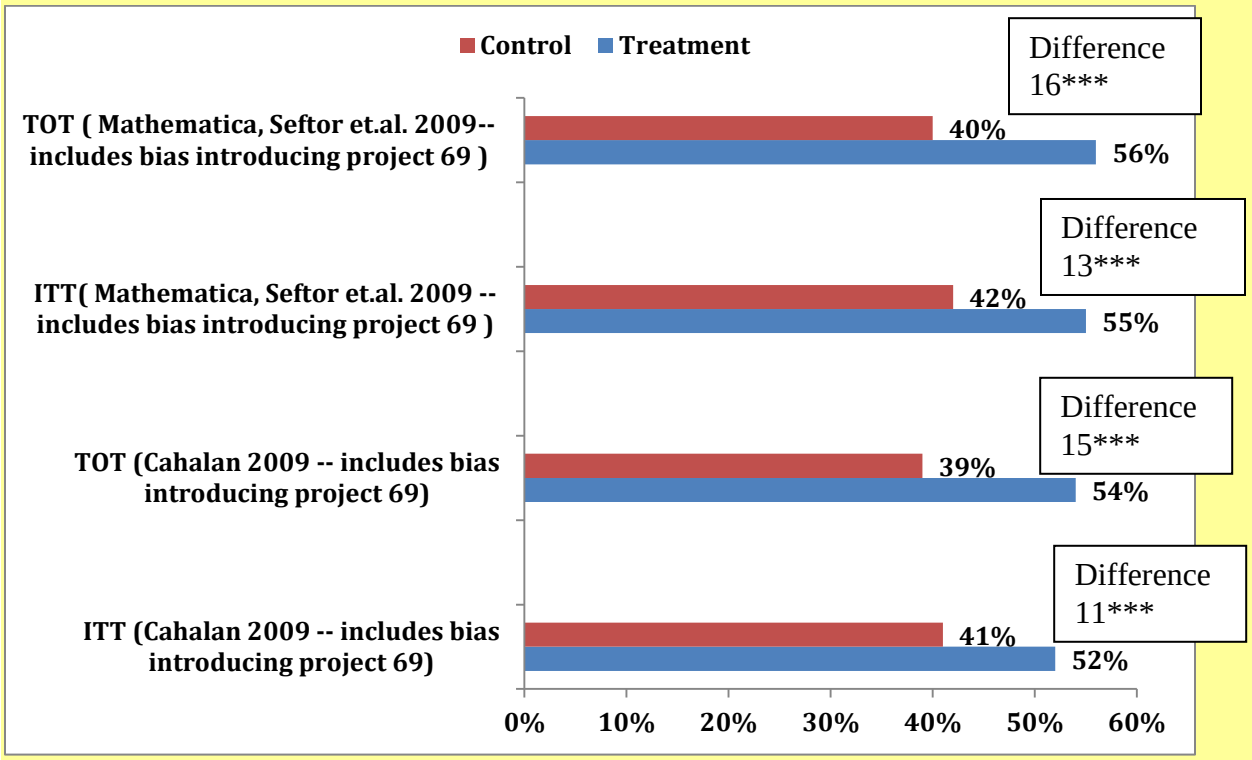
> ***The impacts on receiving any postsecondary credential and receiving a bachelor's degree are 2 and 0 percentage points (effect size = 5 and 0 percent), respectively, and are not statistically significant.***

In making this statement, Mathematica chose to ignore the statistically significant and substantial impacts the study detected on "award of any postsecondary degree or credential" by the end of the study period based on responses to the fifth follow up survey adjusted for non-response without use of the NSC data. These impacts are included in the Mathematica final report in Appendix Tables C7 and C14 (Seftor et.al 2009). These substantial positive findings went unacknowledged in the Mathematica conclusions reported in the body of the report concerning Upward Bound..[7] This data is graphed in Exhibit B-11. As can be seen, Mathematica's own estimate of attainment of "any postsecondary degree or credential" based on responders to the Fifth-Follow-Up Survey shows a positive substantial and significant Intent To Treat (ITT) impact of UB on award of "Any postsecondary degree or credential" of 13 percentage points (55 percent for UB and 42 percent for the control group) and a Treatment On the Treated (TOT) estimate of a 16 percentage point difference. Similar positive UB impacts on award of any postsecondary degree or certificate were reported by Cahalan 2009 available at (see http://www.pellinstitute.org/publications-Do_the_Conclusions_Change_2009.shtml) and are also included in Exhibit B-11. **As noted these large and statistically significant positive impacts, tabulated in their own analyses, were ignored by Mathematica in their text discussion of impact and went unacknowledged in their widely quoted conclusions concerning the "efficacy of Upward Bound". As can be seen above, the Mathematica Executive Summary specifically falsely reports that the study did not detect statistically significant impacts on receiving "any postsecondary credential."** Had Mathematica followed the same procedures as they followed in the 2004 (Third Follow-up) report and the unpublished draft fourth follow-up report basing conclusions on the follow-up survey data adjusted for non-response these would have been the results upon which they based conclusions. However, against PPSS repeated advice to only use NSC data cautiously for BA degree which would have occurred in a later period, as well as that of external IES reviewers to be conservative in use of NSC data Mathematica chose to only put into text tables information with the NSC data that coded non-responders to the fifth follow-up survey without degree information on the NSC as not

[7] Due to positive survey bias, it is possible that the estimates based on survey responders even adjusted for non-response including project 69 overestimates total degree-certificate attainment by the sample, however there were not differences in response rates between the treatment and control group for this round of the surveys—about 75 percent in the 5th follow-up.

having any degree. This violation of NCES coverage standards combined with the bias and instability from project 69's weights and the lack of treatment control group equivalency on academic measures led to a Type II error of failure to report impacts for Upward Bound on award of "any postsecondary degree or credential."

Exhibit B-11: Treatment on the Treated (TOT) and Intent to Treat (ITT) and impact estimates for outcome measure of Award of Any Postsecondary Degree or Certificate by the end of the study period based on 67 of 67 sampled projects respondents to the Fifth Follow-Up Survey



*/**/***/**** Significant at 0.10/0.05/.01/00 level.

NOTE: **Based on 67 of 67 projects sampled**. TOT = Treatment on the Treated; ITT= Intent to Treat **NOTE:** Estimated rates from STATA logistic and instrumental variables regression taking into account the complex sample design. Weighted data use poststratified weights. Cahalan impact estimates used a non-response adjusted weight prepared by Mathematica. Mathematica impacts taken from Appendix Table C-7 and C-14 in the Seftor et.al. 2009 report and are not acknowledged in conclusions reported by Mathematica.

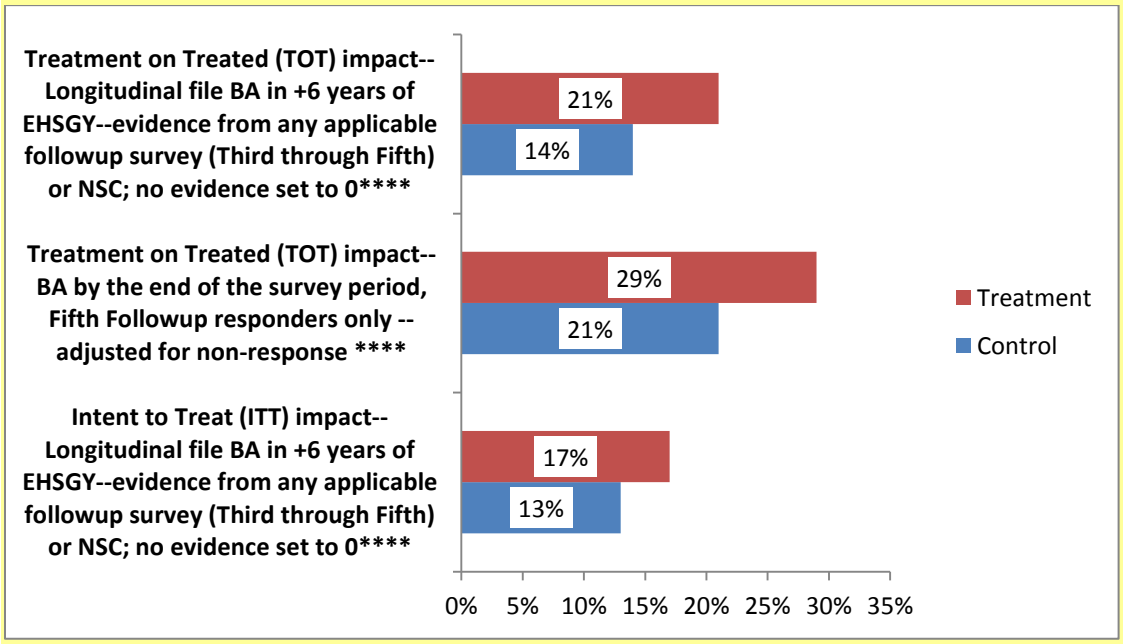
SOURCE: Data tabulated January 2008 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education: study conducted 1992-93 to 2003-04

Violation 8: Failure to Report and Acknowledge Large Positive Impacts on BA Attainment for the 66 of the 67 sampled projects that when taken together had an equally balanced treatment and control group on academic risk factors.

As noted above, Mathematica reported no impact for Upward Bound on BA attainment. Preparation for and fostering BA attainment has historically been a major UB programmatic focus. The UB summer and academic year programs by legislative mandate have a strong academic focus and encourage 4-year

college enrollment whenever possible. Given the representational issues of the heavily weighted project 69 (discussed under violations 1 and 2) combined with the fact that at baseline the treatment group from project 69 was on average on track for certificates or two-year college credentials, and the control group from project 69 was on average on track for advanced degree attainment, (discussed under violations 3 and 4) no impacts were found with or without standardization when the bias introducing project 69 was included in the overall impact estimates. However, PPSS found very strong impacts for the 66 of 67 projects that had a balanced treatment and control group when taken together. None of these more robust estimates for a balanced treatment and control group are acknowledged in the Mathematica reports. Exhibit B-12 summarizes impacts found and reported by the PPSS QA review and sent to Mathematica during the review process for their report. These estimates were ignored in the Mathematica conclusions and reporting that UB had no impact on BA attainment.. For the Treatment on the Treated (TOT) impact PPSS found that UB participation increased BA attainment by 50 percent (from 14 to 21 percent), and for the ITT estimate BA attainment was increased by 28 percent.

Exhibit B-12. Impact of Upward Bound (UB) on Bachelor's (BA) degree attainment: estimates based on 66 of 67 projects in UB sample: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04



*/**/***/**** Significant at 0.10/0.05/.01/00 level.

NOTE: TOT = Treatment on the Treated; ITT= Intent to Treat; EHSGY = Expected High School Graduation Year; NSC = National Student Clearinghouse; Estimates based on 66 of 67 projects in sample representing 74 percent of UB at the time of the study. One project removed due to introducing bias into estimates in favor of the control group and representational issues. Model based estimates based on STATA logistic and instrumental variables regression taking into account the complex sample design. We use a 2-stage instrumental variables regression procedure to control for selection effects for the Treatment on the Treated (TOT) impact estimates. ITT estimates include 14 percent of control group who were in Upward Bound Math Science or UB and 20-26 percent of treatment group who did not enter Upward Bound. Calculated January 2010.

Violation 9: Attribution Error. **Failure to Adequately Acknowledge Issues with Control Group Service Substitution and Treatment Group Waiting List Dropping Out.** *What Works Clearinghouse Standards require that the intervention whose effects are being measured can be clearly attributed to the intervention and that the only difference between the treatment and control group is the intervention. Generally accepted research standards require that the treatment and control group are treated equally except for the treatment; and the treatment and control group are mutually exclusive with regard to the treatment.*

One of the most difficult challenges of education random assignment studies, especially of voluntary support service federal programs, concerns establishing and maintaining clearly distinguished treatment and control groups. This issue has been repeatedly raised by UB study project participants and stakeholders concerning the Mathematica Upward Bound evaluation from the period of the initial random assignment process. This issue also formed the basis of the arguments made in Congress against a new UB evaluation study begun in late 2006, and cancelled by ED in early 2008 following Congressional cutting off of funding. It was argued that it would be unethical to purposively increase recruitment among the 9th graders targeted in the study (those who were most academically deficient and most vulnerable), and then to limit entrance into UB program throughout high school to half of those recruited. If time sensitive services were not denied and alternative services were provided then the study results might be confounded by control group substitution and treatment group dropping out.

This type of confounding was the situation of the Mathematica Upward Bound study. The design was not a randomized control trial in the classic sense—in that the study did not attempt to control, and could not have ethically controlled, student's participating in time sensitive similar pre-college supplemental services from other sources. In the specific two summer window of years, only the opportunity to be invited to apply to the specific regular UB program sampled was done in a randomized manner from a group of middle and early high school students who completed a baseline survey indicating interest in the program. TRIO pre-college programs (Upward Bound, Upward Bound Math Science and Talent Search) in a given area often work together to serve the same target schools with large percentages of students meeting the legislatively mandated eligibility requirements---low-income, disabled, and first generation students. As the data below indicates, a large majority of those students in the control group not selected to be given the regular Upward Bound opportunity, were then given the opportunity for Talent Search, or in some cases Upward Bound Math Science, a new initiative being organized on a regional basis in this period.

The Mathematica Upward Bound study baseline and follow-up surveys contained questions (sometimes quite detailed) about other pre-college support or supplemental service participation, although these questions were somewhat different in each of the applicable survey rounds and are limited by the fact that the students were also in different grades at the time they completed the various survey rounds. They also suffer from the fact that the control group was not asked directly about any regular UB participation.[8] However, sufficient information was collected to classify whether the student reported any other pre-college support or supplemental services, and whether the study participant participated

[8] In asking about alternative services, control group members were given a list of specific and general programs. For the control group the list did not contain the name "UB program." They were asked about UBMS participation.

in UBMS.[9] This information, summarized in Exhibit B-13, can be used to gain some understanding of how much of an issue equivalent and/or similar service receipt was for this study.

Exhibit B-13. Number and percent of study sample participating in UB or UBMS and other pre-college support or supplemental service programs with academic components, by treatment and control group status: National Evaluation of Upward Bound, study conducted 1992-93 to 2003-04

	Random Assigned Treatment		Random Assigned Control		Total Horizons Study	
	Unweighted	**Poststratified Weighted**	**Unweighted**	**Poststratified Weighted**	**Unweighted**	**Poststratified Weighted**
Total	1,524 (100%)	21,866 (100%)	1,320 (100%)	21,866 (100%)	2,844 (100%)	43,732 (100%)
Reported participated in UB or UBMS service	1,247 (82%)	17,843 (82%)	180 (14%)	2702 (12%)	1,427 (50%)	20,545 (47%)
Reported participated in "another" (not UB and not UBMS) pre-college support or supplemental service program only	128 (8%)	2,332 (11%)	618 (47%)	10,513 (48%)	746(26%)	12,845 (29%)
Did not report participation in any type of (UB, UBMS, or other) pre-college support or supplemental service program	149 (10%)	1690 (8%)	522 (40%)	8651 (40%)	671 (24%)	10,342 (24%)
Reported participated in any type (UB, UBMS, or other) of pre-college support or supplemental service program	1375 (90%)	20,176 (92%)	798 (61%)	13,215 (60%)	2173 (76%)	33,390 (76%)

NOTE: Percentages given in parentheses. UB = Upward Bound; UBMS = Upward Bound Math/Science. Weighted data use poststratified weights for longitudinal file. **SOURCE:** Data tabulated January 2008 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education: study conducted 1992-93 to 2003-04; and federal Student Financial Aid (SFA) files 1994-95 to 2003-04.

[9] Information was collected on the surveys about length of participation and type of programs on the various surveys that could be analyzed in more detail.

Participation in the UB Program or Upward Bound Math Science by the Treatment and Control Group. About 26 percent of the Treatment Group maintained in the ITT analysis which formed the basis of Mathematica's conclusions was coded as "waiting list" dropouts during the period in which cases were randomly selected to be recruited to apply for entrance into the program, and about 20 percent of the treatment sample reported on the First Follow-up that they did not enter regular Upward Bound.[10] A portion of this 20 percent reported they could not remember being given the opportunity when asked about it a year later. Conversely about 12 to 14 percent of the control group reported they entered into Upward Bound Math Science (UBMS) or Upward Bound. The Mathematica Fifth-Follow-up Report, while emphasizing Intent to Treat (ITT), includes some Treatment on the Treated (TOT) analysis taking into account the 12-14 percent UBMS crossovers and the treatment non-UB participants (unlike the Third Follow-up report which did not recognize UBMS participation by the control group as a crossover). However as one of the IES external reviewers noted given that one fourth to one-fifth of the so called "treatment group" did not enter UB and 12-14 percent of the control group was in UBMS or UB, the TOT estimates may be more meaningful statistic for this study. In the study year procedures were altered to ensure there would be double the number of baseline survey completers as openings. As noted, instead of obtaining actual applications to UB, those who completed the baseline surveys were considered on a "waiting list" for participation and in the study years no one could get on the UB "waiting list" without completing the baseline survey. All of the students were minors and over half were in middle school when completing the baseline survey; hence their actual entry into the UB program that next summer which was typically a residential program was related to parental permissions and family mobility. Low income families have high levels of mobility.

Participation in other lessor pre-college services. Examination of study survey data also revealed that a majority (60 percent) of the control group reported participating in some form of supplemental pre-college programs by the end of high school (Exhibit B-13). Frequently when students were not randomly selected for Upward Bound, they were placed in Talent Search or another similar pre-college program. Presumably most of these programs were less intensive than Upward Bound. Cahalan reports that PPSS requested that Mathematica use the information from the baseline and follow-up surveys on alternative service receipt to statistically address issues of service substitution and treatment group non-entry into UB; however, Mathematica declined to conduct these analyses.

This is the same issue addressed by noble laureate James Heckman, and co-authors (Heckman Hohman, Smith, and Khoo 2000) re-analysis of the Job Training Partnership Act (JTPA) evaluation in which they considered the interpretation of evidence from social experiments when persons randomized out of a program being evaluated have good substitutes for it, and when persons randomized into a program do

[10] There was also a group of about 192 "waiting list participants" who completed baseline surveys who were excluded from the analyses and their weights re-distributed among those randomly assigned because they were selected with certainty into the group being given the UB opportunity. Projects were allowed to serve selected students in Upward Bound in the study period to whom they might have made a prior commitment or if the projects wished to serve them for group cohesion or diversity purposes.

not enter the program or drop out. Using data from an experimental evaluation of JTPA classroom-training programs, they documented the empirical importance of control group substitution and treatment group dropping out. They note that "evidence that one program is ineffective relative to close substitutes is not evidence that the type of service provided by all of the programs is ineffective, although that is the way experimental evidence is often interpreted" (Heckman et. al. 2000).

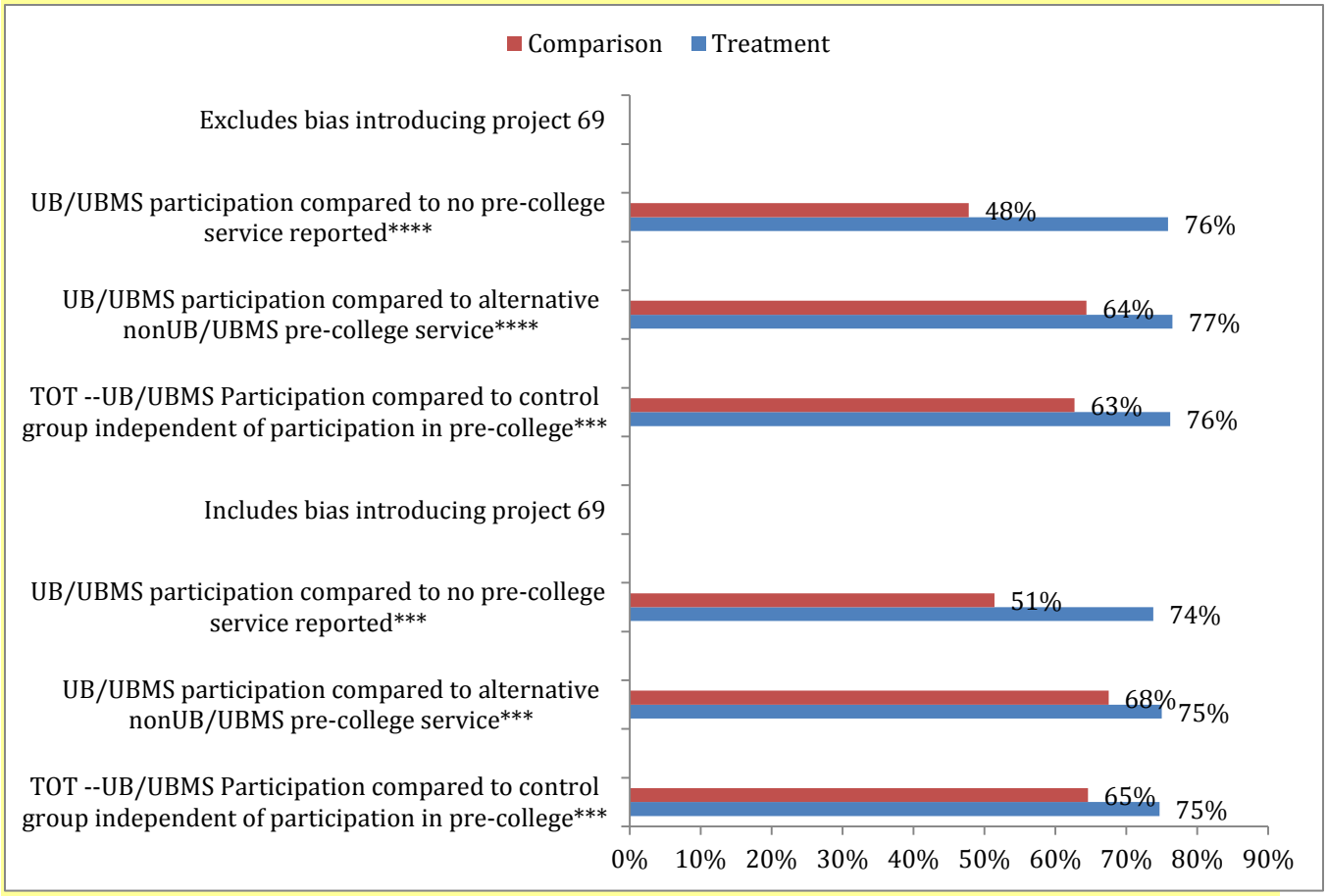
None of the Mathematica reports include comparisons of UB participants with those getting only other services or a serious consideration of the contamination issues related to the receipt of other similar but less intensive non-UB/non-UBMS services that were received by a majority of the control group. Cahalan reports that concern with the serious contamination issues, led one internal PPSS reviewer of the Mathematica Fifth Follow-up Report, Dr. Jay Noell, the PPSS UB Evaluation Technical Monitor/COR prior to Dr. Cahalan, to recommend that Mathematica acknowledge that the random assignment study had too high a level of contamination to be valid due to these similar but less intensive alternative services that had been provided often—precisely because the students did not get randomly chosen for Upward Bound. He recommended that the data be analyzed as a quasi-experimental design using instrumental variables regressions to model factors related to different levels of participation and to use these estimates in the second stage to control for the observed selection differences with regard to type of participation. Dr. Noell recommended that the ITT analyses be placed in an appendix, but not be used in assessing program effectiveness.

The *Cahalan Re-Analysis Report* presents the ITT and TOT analyses using models and methods similar to Mathematica except for standardization of outcomes and avoiding use of NSC for non-respondent enrollment and below BA degree estimation, but, following Dr. Noell's recommendation also includes some additional observational quasi-experimental design analysis using instrumental variables regression. These analyses compare outcomes for those who were in UB/UBMS with those who reported they participated in some other non-UB/UBMS pre-college supplemental service (see *Cahalan Re-analysis Report* chapter 4) and with those who reported not participating in any pre-college supplemental service program.

Exhibit B-14 gives results for postsecondary entrance in +4 years after expected high school graduation year and Exhibit B-15 gives results for award of BA degree in +6 years after expected high school graduation year. Both Exhibits show strong impacts for Upward Bound compared to both those who reported participation in some alternative less intensive pre-college service and compared to those who reported not participating in any pre-college services. For example, when **appropriate analyses controlling for selection bias are conducted, UB participants were 3.3 times more likely to obtain a BA in 6 years compared to those with no pre-college access services and 1.4 times as likely as those participating in only a less intensive service program such as Talent Search.** These strong positive results are not acknowledged in the reports. This failure to address this issue in the Mathematica reports has led to other researchers (see Haskins and Rouse 2013) to mistakenly

generalize from the erroneous Mathematica reports, that conclude that all federal college access programs have been shown not to be "effective". In fact 75 percent of the entire sample (treatment and control) had some form of supplemental pre-college services, and typically this was from another federally supported program.

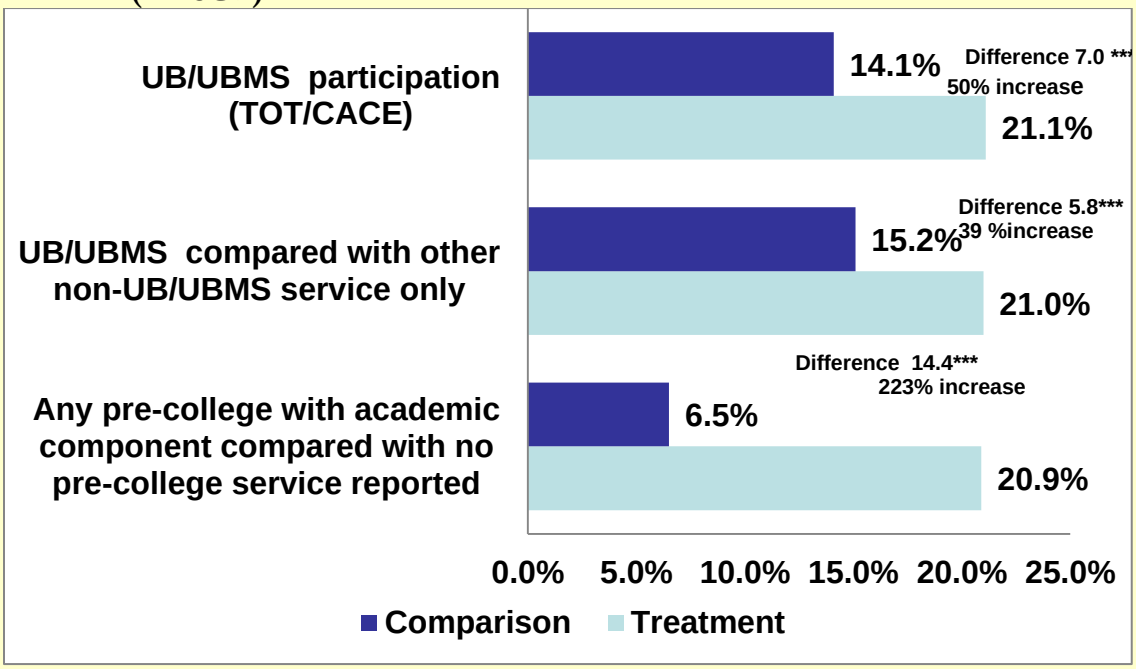
Exhibit B-14: Estimates of relative impact of participation in various levels of pre-college access supplemental services on entry into postsecondary education within +4 years after expected high school graduation: National Evaluation of Upward Bound



*/**/***/**** Significant at 0.10/0.05/.01/00 level.

NOTE: Based on 66 of 67 projects sampled. The estimates in the figures shown are based on longitudinal data over a 10- year period in an analysis using instrumental two-stage regressions that first model factors related to differences in participation in services and then use these factors in the second stage to control for participation selection bias factors.
SOURCE: Cahalan, Margaret: *Addressing Study Error in the Random Assignment National Evaluation of Upward Bound: Do the Conclusions Change?* The report can be accessed at the following site: http://www.pellinstitute.org/publications-Do_the_Conclusions_Change_2009.shtml

Exhibit B-15 . Instrumental variable Regression Results from the National Evaluation of Upward Bound for BA attainment in +6 years after Expected High School Graduation Year (EHSGY)



*/**/***/**** Significant at 0.10/0.05/. 01/00 level.

NOTE: TOT = Treatment on the Treated (TOT); UB = Upward Bound; UBMS = Upward Bound Math Science. All estimates significant at the .01 level or higher. Estimates based on 66 of 67 projects in sample representing 74 percent of UB at the time of the study. One project removed due to introducing bias into estimates and representational issues. We use a 2-stage instrumental variables regression procedure to control for selection effects. SOURCE: Data tabulated January 2010 using: National Evaluation of Upward Bound data files, study sponsored by the Policy and Program Studies Services (PPSS), of the Office of Planning, Evaluation and Policy Development (OPEPD), U.S. Department of Education; study conducted 1992-9 to -2003-04.

Violation 10: **The Mathematica UB reports violate both the Joint Committee for Educational Evaluation Proprietary Standards that Stakeholders be fully informed concerning data that affects them and the AERA Standards related to the "sufficiency of the warrants and the transparency of the report"** *The Joint Committee Education Evaluation Standards state:*

> *P-6. The formal parties to an evaluation should ensure that the full set of evaluation findings along with pertinent limitations are made accessible to the persons affected by the evaluation and any others with expressed legal rights to receive the results.*

This is among the most serious of the violations with the Mathematica reports. They are non-transparent in reporting positive impacts detected by the study and also in describing study issues—stating that the heavily weighted project 69 was typical of its stratum, and did not make a difference in conclusions. The reports also are not transparent in reporting the treatment-control group lack of balance on academic risk factors. The reports also state that the heavily weighted project 69 driving their no-impact

conclusions had "below average performance" when in fact the significant negative impacts found when this project is considered alone ---were due to the extreme uncontrolled differences between the treatment and control group in this project showing a failure of the random assignment procedures in this site. The reports are especially troubling because not only do they contain erroneous conclusions, but they also contain seemingly deliberate statements that mislead readers into thinking that the legitimate concerns, identified by PPSS internal and external reviewers, had been adequately addressed and did not make a difference in the study conclusions. **The QA re-analysis, however, conducted by Ed-PPSS technical monitoring staff found the assertions by Mathematica Policy Research of "no detectable impact" except for the award of certificates to be false.**

Detailed Reports Documenting Standards Violations and Re-Analysis Results

Detail on the major issues with the Mathematica reports and results of standards based re-analyses have been presented in several documents publically available at the addresses noted below.

- *Addressing Study Error in the Random Assignment National Evaluation of Upward Bound: Do the Conclusions Change? By Margaret Cahalan* a COE report published in 2009 and available at http://www.pellinstitute.org/publications-Do_the_Conclusions_Change_2009.shtml.
- *The Council for Opportunity in Education (COE) Request for Correction submitted in 2012* is available at http://www.coenet.us/files/pubs_reports-COE_Request_for_Correction_011712.pdf,
- *Expert Statement of Concern* with Regard to the Mathematica National Evaluation of Upward Bound signed by leading researchers can be found at http://www.coenet.us/files/ED-Statement_of_Concern_011712.pdf. The Statement accompanied the above referenced RCOE Request for Correction and was signed by researchers who had reviewed the request and found it cause of serious concern. The signers to the UB Evaluation Statement of Concern included the sitting presidents of the American Education Research Association (AERA) and the American Evaluation Association (AEA).
- Attachment D included with this package—Draft *Flawed Reports from the National Evaluation of Upward Bound Masked Significant and Substantial Positive Impacts: The Technical Monitors' Perspective* by Cahalan and Goodwin, forthcoming 2014