EVALUATION ERRORS: SIXTEEN LESSONS FOR RESEARCHERS AND PRACTITIONERS FROM THE MATHEMATICA UPWARD BOUND STUDY

BY MARGARET CAHALAN

Abstract

The long running National Evaluation of Upward Bound (1992-2009) was considered a gold standard random assignment study designed to assess the average impact of the Upward Bound Program. After the study had been active for over a decade, and published several impact reports, a quality assurance review conducted by ED found serious sample design, treatment control group bias and analyses and reporting issues that were serious enough to impact the study conclusions.

Using the Upward Bound evaluation as a case study, this essay presents 16 lessons learned for researchers and practitioners alike for future evaluations in designing, implementing, analyzing, and reporting results.

The Trump administration's proposals to continue funding only those programs that have "evidence of effectiveness" and the associated President's FY2018 budget proposals to cut critical service programs, including TRIO and GEAR UP programs, have prompted me to prepare this essay. Mr. Trump's threats and proposals are similar to those put forth by President Bush in his FY2005 and FY2006 budgets calling for zero funding of Upward Bound (UB), Upward Bound Math Science (UBMS), Talent Search (TS), and GEAR UP. These proposals were justified in President Bush's budgets by the early and interim findings from the Mathematica Policy Research's (Mathematica) National Evaluation of UB (Myers et.al. 1999; Myers et.al. 2004) that were later found to be erroneous (Cahalan 2009; Cahalan and Goodwin 2014; Nathan 2013; Harris et.al. 2014). It was only after extensive practitioner lobbying in Congress in 2006, by COE and other college opportunity groups that the programs were ultimately saved, albeit with continual decreases in number of students served due to the resulting level funding for over a decade. This level funding was only checked after 2016 with increases for TRIO in the 2017 and 2018 budgets.

This essay is also written in response to the May 2017 blog posting by Paul Decker, CEO of Mathematica Policy Research (Mathematica). Dr. Decker's 2017 blog posting laments the lack of using research for program improvement rather than just the "cutting block." Few would disagree with Dr. Decker's point that research should be used for program improvement; however, his Blog posting once again uses Mathematica's erroneous UB evaluation results as an example of what research can contribute to program improvement. The

Reprinted from *Reflections on connecting research and practice in college access and success programs*, pp.21-34, by H. T. Rowan-Kenyon, M. Cahalan, & M. Yamashita, Eds., 2018,Washington, DC: Pell Institute.

article contains no acknowledgement of the errors in the Mathematica reports. Nor does it argue for more valid and improved research. It also does not acknowledge the limited level of college preparation services available for low-income and first-generation students. Rather, the article argues for more funding for research of the same type as conducted by his company, Mathematica. Dr. Decker's 2017 blog posting is a shorter, less aggressive form of his 2013 APPAM Presidential address (Decker 2013) but again omits any mention of the glaring errors revealed in a Quality Assurance Review conducted by the technical monitors of the final of the three Mathematica Upward Bound evaluation contracts within the U.S. Department of Education (ED).

As a lifelong contract researcher as well as a Technical Monitor for evaluation studies, I wish to share with you the specific evaluation lessons that I believe can be learned from my unsettling experience when I became employed within the U.S. Department of Education (ED) and served as the supervisor of a group responsible for the technical monitoring of a number of postsecondary evaluations. These studies included the final Mathematica contract for the long-running National Evaluation of Upward Bound. In this essay, I too, as did Dr. Decker, use the Mathematica Upward Bound study as an example. However, it is an example of the pitfalls that can happen to an evaluation when practitioner and stakeholder rights to a valid and transparent evaluation are not followed (see Joint Committee for Education Evaluation Standards, 2011). These lessons learned will hopefully inform both practitioners and researchers in the design, implementation, and analyses of future studies, in an era in which federal service programs are in serious jeopardy.

The details of the errors in the Mathematica report, along with fully-documented positive and substantial impacts found when these errors are addressed using NCES statistical standards, are reported in documents published by the Council for Opportunity in Education (COE) in 2009 and 2014 and in publications of researchers from the University of Wisconsin and Tulane University (Cahalan, 2009; Nathan, 2013; Cahalan and Goodwin, 2014; Harris, Nathan, Marksteiner, 2014). Links to these reports, and to COE's 2012 formal *Request for Correction* can be found at the following site http://www.pellinstitute.org/publications-Upward_Bound_Compilation_of_Links.shtml.

Background Overview of the Study. The National Evaluation of Upward Bound was hailed as a gold standard longitudinal random assignment study when it was begun in the early 1990s. The study was designed to follow about 3,000 middle school and early high school students interested in Upward Bound, who were randomly selected to be either given the opportunity for UB participation in the study years or not (treatment or control group). The design called for following both groups through at least six years after scheduled high school graduation. The RFP for the study called for a nationally-representative random assignment study design that would allow for estimating the average national impact of the Upward Bound program. The study design also included a descriptive Project Survey and several qualitative case studies of projects to gain an overall picture of UB services provided. Projects were also asked to keep detailed service records for study participants.

Below, as I unfold the UB evaluation story, I list 16 interrelated lessons learned from the Mathematica National Evaluation of Upward Bound. While this essay is especially addressed to

researchers, it is also addressed to practitioners. I offer it in the hopes that the lessons learned will lead to new models for evaluation that emphasize respectful partnerships between practitioners and researchers.

Sixteen Lessons Learned From the Mathematica Upward Bound National Evaluation Story

1. ***Consult with All Stakeholders and Take Seriously the Feasibility Concerns of Practitioners Concerning What Will and Will Not be Feasible and Ethical.*** After the basic UB evaluation design had already been developed and at the time the study was first presented to the Upward Bound community, a number of concerns were raised by program practitioners who were required to participate. Practitioners expressed concern that the study would change the manner in which they interacted with potential participants, especially in recruitment, and also questioned whether the national study would yield valid and useful information for their differing local projects. A major concern had to do with the anticipated difficulty of actually implementing a random assignment study that would not result in IRB- prohibited denial of services concerns. There were also concerns expressed about the potential for control group contrast. Projects did not like the changes to the normal recruitment procedures they followed and specifically did not like the requirement that in the study years they needed to recruit at least twice the number of students than the anticipated openings and then purposely not give UB services to half of the students recruited. While project stakeholders were informed of the study requirements before implementation, they were not consulted on the basic study design. In large part these concerns by practitioners were ignored, and as will be seen below the accommodations made for stakeholders seem to have made the design even more problematic.

2. ***Do Not Over-Promise and Sacrifice What Might be Possible for That Which Is Not Possible.*** In order to make the study seem more useful to stakeholders, as the study planning proceeded, the sample design evolved to attempt to allow for estimating not just national impacts but the differences in impacts between different types of projects hosted at different types of colleges (2- or 4-year); public or private; rural or urban; and serving participants of different race/ethnicities; and projects of different sizes. It was this attempt to have both national estimates and estimates for many different types of projects that resulted in a highly-stratified first stage sample of 67 projects drawn from 47 strata, some of which had only one project. Moreover, due to the need to respect projects' existing patterns of recruitment, the final second stage student level sample was drawn from over 300 strata. In the final stage of inverse of the probability of selection weighting, one project, the single representative of a very large stratum, ended up carrying fully 26 percent of the weights. As we will see, as the UB evaluation story unfolds, the project sample design sacrificed possible overall estimates for a poor design incapable of producing either reliable national estimates or separate estimates for the various strata, and with serious unequal weighting.

3. ***Do Not Assume an Error-Free Study Implementation of a Random Assignment Design.*** The contractor, Mathematica, is known for its expertise in "random assignment" and presented the procedures of the UB evaluation as the gold standard. In fact, there are

very few studies that attempt weighting using inverse of probability of selection for a multi-stages stratified sample with a random assignment design. As noted above, projects were instructed to carry out recruitment in a special way in the study period. This consisted in recruiting at least twice the number of students as anticipated openings to be on what was called a "waiting list" for Upward Bound. Only students who completed a detailed "baseline survey" were allowed to be eligible for the "waiting list" to apply for openings in the Upward Bound program over the two recruitment study years. However, projects implemented the request to submit at least twice the number of baseline surveys completed by potential participants as anticipated openings in different ways. Some administered the baseline surveys to large numbers of students who might have been in Talent Search (another TRIO pre-college program often beginning in middle school) or recruited from a whole 8th or 9th grade class in a low-income target school. Other projects submitted far fewer baseline surveys from students who might have been closer in time to being what might be considered "formal applicants." As the weighting at each stage was the inverse of the probability of selection and the second stage weighting was done relative to the number of baseline surveys submitted rather than the actual number of openings, this difference in the number of baseline surveys relative to actual openings contributed further to the unequal weighting and balance issues. To feasibly implement the random assignment stage at the UB project level, the different projects were allowed to establish their own project-specific strata. For example, they could establish separate strata with different probabilities of selection by gender, or for the different target high schools served. This involved dividing the so-called "waiting list" by the additional stratification criteria —so in the final combined stage, as noted there were over 300 strata. Within a given project students thus had different probabilities of selection and second stage weights. The control group consisted of all baseline survey completers who were not randomly selected to be given the UB opportunity as openings developed within the project strata. Thus, the control group and treatment group did not have equal actual numbers within a given project. Weight adjustments were used to equalize the final stage weights so that the control group and treatment group in each project has an equal sum of weights. In addition, about 10 percent of the baseline survey submitters were selected with certainty to be in the UB program, due to previous commitments or group dynamics reasons and were removed from the study with their weights redistributed among the study treatment cases in their strata.

4. ***Given Denial of Services Concerns in College Access, Do Not Assume that the Control Group in Practice Will Not Get Related Services.*** As openings for the UB program occurred over two summers, within the project strata, students were selected at random to be given the opportunity for program participation. Those not randomly selected became the control group. Subsequently, given student mobility and other factors, it was found through surveys and project records that about 25 percent of those selected to be given the UB opportunity did not report ever entering into Upward Bound, with some portion of these reporting they did not remember ever being given the opportunity. These students were considered in the UB Treatment Group in the "Intent to Treat (ITT)" analyses. Conversely, about 15 percent of the Control Group entered into Upward Bound or Upward Bound Math Science and were maintained in the Control Group. In addition and importantly, over a majority (60 percent) of the Control Group reported getting some form of other supplemental pre-college services in high school; most frequently, it was the less intensive Talent Search

(another TRIO program). Some UB practitioner staff reported that, in order not to be denying time-sensitive college access services, they actually tried to find alternative pre-college services for those in the control group who did not get randomly selected for the Upward Bound opportunity.

5. ***Reveal Any Issues Early—the Importance of Transparency.*** By the time I joined ED, three Mathematica reports had been published in 1996, 1999 and 2004. It was not until 2005, soon after I joined ED and after the study had been active since 1992, that our ED-PPSS division director, Dr. Goodwin, and I were alerted to some problems with the UB study. This alert first came from a new Mathematica staff person working on analyses and report writing of the Fourth Follow-up, who subsequently left the company. The staff person was concerned after having found that the "no overall impact results" were being driven by only one project of the 67—with the unusually large weight, accounting for 26 percent of the total sum of the weights. It was further reported to us that this project, when considered alone, had an unusually large number of applicants (baseline surveys submitted) and also had very large negative impacts. These negative impacts were strong enough to impact the overall positive impacts found when this project was omitted. Because this project was supposedly representing a very large number of both projects and baseline survey completers (considered applicants), this flawed design meant that the outcomes of some students from this project "waiting list" carried weights that were 40 times those of the lowest weighted students from other projects. The first three Mathematica UB evaluation reports, published over almost a 10-year period, did not reveal these serious sample design issues.

6. ***Investigate Data Results that Defy Logic.*** At the time, as ED Technical Monitors, we asked Mathematica to investigate why the heavily weighted project was having such large negative impacts on college going so that we could understand what the project was "doing wrong." We could understand "no impact" results, but very large significant negative impacts did not seem plausible. Unfortunately, we did not get a response from Mathematica on this request. Instead we received a draft of the Fourth Follow-up report that, similar to the three previous Mathematica reports, did not reveal the unequal weighting issues, or that the lack of impact results were being driven by only one project. In the course of reviewing the Fourth Follow Up report as Technical Monitors, we were also struck by the very large positive significant impacts found for the sub-group of students that were classified as being of high academic risk at the start of the study. The results reported that there was an over 20 percentage points significant positive UB impact on college enrollment based on differences between the academically at-risk group treatment and the at-risk control group in college entrance. At the same time, it was reported that there were no overall impacts. We were also struck that, contrary to NCES statistical standards, the outcome measures had not been standardized by high school graduation year for a sample that spanned five years of scheduled high school graduation dates (given that the study recruited students from 8th grade to 11 grade and over two summer program rounds). When we asked Mathematica to standardize the outcome measures, Mathematica responded that they did not need to do this standardization given that there had been a random assignment that should have resulted in equal division between the treatment and control group on this distribution. However, ED monitoring staff balance checks later found that the control group was on average in a higher grade at the start of the study than the treatment group and that when we standardized

outcomes that the overall conclusions changed—even with the outlier heavily weighted project with 26 percent of the weight and negative impacts.

7. ***Pay Attention to Quality Review Results.*** Given the lack of response of Mathematica to our concerns, in 2006, as Technical Monitors, we requested the data files from Mathematica and began an internal ED-PPSS Quality Assurance (QA) review of all data files from the study. At first, we were primarily concerned that the study was relying solely on survey results and that the non-response issues were aggravated by the unequal weighting. Knowing that the administrative records from the federal student aid application and award data was available for the entire sample, we arranged to have the sample matched to the federal aid files and merged on to the UB study files for analyses. We also consulted other external statistical contractors and sent the Upward Bound data files to RTI, which held the ED-PPSS Statistical Technical Assistance contract at the time, for review and replication of our internal findings. The ED staff internal and RTI external quality reviews revealed that the study had serious sampling and non-sampling error issues that had not been revealed by Mathematica over the course of the more than a decade of contracts and after the publication of a major impact report in 2004. On the basis of using the federal aid files, ED-PPSS internal analyses, replicated by RTI, began to find positive overall impacts on postsecondary outcomes.

8. ***Make Sure there is an Actual Balance between the Treatment and Control Group through Balance Checks.*** The balance analysis done as part of the Quality Assurance Checks revealed that the negative impacts in the outlier weighted project reflected a failure of the implementation of the multi-stage random assignment rather than negative project impacts. The reason for random assignment is to ensure an equivalent treatment and control group on factors associated with the outcome or a balanced treatment and control group. The negative impacts observed in the project carrying 26 percent of the weights were due not to negative practices but to large differences between the treatment and control group on academic risk, educational expectations, and grade at baseline within this project. In what can only be a failed implementation of the random assignment, 80 percent of the high academic risk students from this project were in the treatment group and 20 percent in the control group. The control group from this project was also on average in a higher grade in high school at the start of the study. In fact, the treatment group from this project was contributing fully 30 percent of the overall weight for students of high academic risk in the overall sample, while the control group from this project was contributing a larger proportion of the students that were academically talented and had advanced degree expectations. Among the total sample, an average of 36 percent expected an advanced degree, compared with 56 percent in the control group from the outlier project, and 15 percent of the treatment group from this project. This explained the large positive impacts for the high academic risk sub-group in the analyses. When only those with high academic risk were considered, the high-performing, highly-weighted members of the control group from the unbalanced project was not included. Within the high academic risk sub-group, there was by definition a more balanced treatment and control group. Hence the high UB sub-group positive results for at-risk students emerged (Cahalan 2009).

The large weights among this unbalanced project combined with the lack of treatment and control group balance led to an unacknowledged and uncontrolled bias in favor of the control group among the overall sample forming the basis of all of the published Mathematica national estimates. This issue combined with the lack of standardization of outcome measures by high school graduation year and other analysis and reporting errors were serious enough to impact the overall conclusions Mathematica reported concerning the Upward Bound project. Moreover, (contrary to what Mathematica had reported in 2004, continued to report in 2009, and was repeated by Dr. Decker in his 2013 APPAM Presidential Address) strong positive impacts for the overall sample, including the bias-introducing outlier project, were found for UB when these errors were addressed by such common statistical standards requirements as standardizing the postsecondary outcome measures by scheduled high school graduation date.

9. ***Check the First-Stage Sample for Atypical Cases that Cannot Represent Their Strata before Proceeding and Check the Eligibility of those Considered "Applicants".*** To protect confidentiality, the data files delivered to ED in 2006 and 2007 for QA review did not reveal the identity of the UB projects in the sample. In the course of reviewing the study, Dr. James Chromy, the RTI statistician we consulted, asked us to request a copy of the sampling frame from Mathematica in order to gain a better understanding of the sample design characteristics. We made this request of Mathematica in early 2007; however, Mathematica indicated that they did not have an electronic version of the sampling frame and could not locate the paper version for nine months. Consequently, this frame was not delivered to ED until a few weeks before the final contract ended in late 2007. Until this time, no one at ED or RTI was aware of the identity of any of the study projects. A review of the sampling frame in late 2007, (something that should have been done by Mathematica in the 1990s, before the study began), revealed significant facts that explained the strangeness of the results we were observing. In a flawed design, the project carrying 26 percent of the weight was selected to be the **single** project representing the UB projects hosted at 4-year public, urban, non-majority Hispanic institutions, and average- sized UB project. In other words, this single project was representing a lot of UB projects. This, combined with the large number of baseline surveys submitted, resulted in its very large outlier weights. To compound this problem, when we researched this project's characteristics, we found that unfortunately it was very atypical for the 4-year public stratum for which it was supposedly the sole representative. It was housed at a former private junior college serving minority youth from minority vocational high schools that historically awarded postsecondary trade certificates. The junior college had been taken over by a large public city college system and hence its 4-year public formal classification in IPEDS. *This accounted for the fact that the only positive overall impacts that Mathematica reported were for the award of certificates*. Dr. Chromy and the other external reviewers noted that having only one atypical project representing the largest stratum of 4-year bachelor's degree granting projects could not produce robust estimates for that stratum or for the whole, and that this was especially problematic for any estimates of bachelor's degree attainment. Checks should have been done of the design, and on the actual randomly drawn first stage project sample prior to beginning the study to ensure that the cases drawn were not "atypical" for the strata they were representing. In this case, Dr. Chromy noted that the sample should have been re-drawn prior to beginning the study.

The representation issues of this heavily-weighted project were compounded by the actual non-equivalence of the treatment and control group in this project. For unclear reasons, the implementation of the random assignment in this outlier project appears to have been flawed. The Treatment Group resembled on average the types of vocational students historically served by this regular UB program, while the control group on average resembled the types of students most frequently served by Upward Bound Math Science (UBMS)--in a higher grade at entrance, more academically talented, and with higher advanced degree expectations at baseline. Indeed it is unclear whether these control group members were actually completing the "baseline survey" because they were interested in the sampled particular Regular Upward Bound program with ties to vocational certificates, or actually had more of an interest in a new UBMS program in the region, (not in the sample) in which some of the control group from this project reported participating. These "baseline survey completers" were kept in the control group and were probably not serious applicants for the (at that time) vocational certificate-focused UB project.

10. ***Understand that Response and Coverage Issues Are Important; Follow NCES Standards for Response and Coverage; Note that Unequal Weighting Can Make the Non-Response Issues More Significant.*** One of the most problematic aspects of the Mathematica analyses in the final 5th follow-up 2009 report was the ignoring of NCES coverage standards and the use of the National Student Clearinghouse (NSC) at a time when the coverage by institutions reporting to the NSC was about 27 percent for enrollment and when NSC was not yet collecting degree information. This coverage issue was especially problematic for less than bachelor's degree-granting postsecondary institutions and was far less than would meet NCES standards for coverage. In the 5th and final report, survey non responders who were not found on the NSC data base were assumed to have not entered into postsecondary education or not to have obtained a postsecondary credential. This issue was aggravated by the unequal weights. For example, some sample members had weights of 158, whereas the lowest-weighted sample member carried a weight of 4. This was especially problematic as the heavily-weighted outlier project was not reporting to the NSC in the most applicable period when the students would be entering and completing postsecondary credentials.

11. ***Make Sure that the Conduct of the Study Is Integrated.*** The Mathematica formal structure provides for separate groups responsible for various aspects of a study, the statistical staff, data collection staff, qualitative case study staff, and econometric staff. Given the very long time period covered by the study there was also considerable turnover of staff. It is therefore not known if the project leadership or analysts of the various groups knew that the highly-weighted, vocationally-focused UB project was supposed to be the only representative of a large 4-year public hosted stratum, or that the treatment and control group were so unbalanced in this site. It is known that the separate significant "negative" impacts of this site were observed by 2005. We also do not know how many of the issues revealed in the ED-QA review were known to junior and mid-level Mathematica staff working on the study, at various times over the 15-year study. It is not known why the study error issues and the unequal weighting were not included in the reports through the third follow up in 2004. It was not until 2005/06 that these issues were called to ED's attention and the full extent of the issues was not revealed at this time. The seriousness of the errors became known by ED

only gradually as ED-PPSS and external reviewers examined the data files and the sampling frame for themselves between 2006 and 2008. The federal aid matches were done in 2006 and as noted the sampling frame was not delivered to ED-PPSS until late 2007 when the contract was essentially over. It was not until the contract was officially over in 2007- 2008 that ED-PPSS internal staff worked to standardize results by high school graduation year and with this mitigation found substantial and significant positive impacts with and without the outlier project in college entrance, award of federal aid, award of any postsecondary degree, but (as might be expected given the unequal composition of the treatment and control group) not for bachelor's degree receipt. The outlier project was not representative of its 4-year stratum and had seriously unequal treatment and control groups that introduced bias in favor of the control group for bachelor's degree receipt that given the large weights was large enough to override the clear large positive impacts on bachelor's receipt when the unrepresentative and bias introducing project was removed from the analysis. When the outlier project was included there were large impacts on the award of any postsecondary credential due to the large impacts for certificate receipt.

12. ***Respect Stakeholder Rights to a Transparent Evaluation with Warranted Conclusions.*** The results of the ED-PPSS technical monitor's analyses were fully shared with Mathematica over the period of 2006 to 2009 as they became known—however, they were repeatedly disregarded by Mathematica project staff. Mathematica's leadership in formal letters accused the ED-PPSS technical monitors as overstepping and acting as advocates for Upward Bound. Memos to Mathematica concerning the data quality review results and errors identified written by myself and also by Dr. David Goodwin, the original UB study project monitor and the head of the ED-PPSS Division responsible for the study in 2008, were repeatedly disregarded by Mathematica project staff and leadership. The repeated failure of Mathematica to reveal the issues with the study to ED and the stakeholders over a period of more than a decade, and their failure, to this current time in 2018 to acknowledge the results of the PPSS QA analyses finding positive impacts, constitutes a serious negligence of the trust of that must be present in all evaluation contracts. This behavior also constituted a serious violation of trust to the Upward Bound and TRIO community. Stakeholders to evaluations have a right to a transparent and ethical evaluation in which only warranted conclusions are put forth concerning their effectiveness.

13. ***Acknowledge All Findings, Especially if They Have Different Conclusions in Transparent Reporting.*** The statistically significant and educationally meaningful positive results PPSS internal staff and RTI external quality review consultants had found, and which had been fully shared with Mathematica by 2008, went unacknowledged in the 2009 published Mathematica final report. When ED staff members applied NCES standards to the analyses (such as standardizing outcome measures, and respecting coverage standards for use of the NSC data) significant and substantial Intent to Treat (ITT) and Treatment on the Treated (TOT) positive impacts were observed for the entire sample, with larger impacts when the problematic bias introducing non-equivalent treatment and control group project was included. The final Mathematica report ignored their own analyses (presented in an appendix to their final report) that found a substantial and significant 12 percentage point-effect size for the award of any postsecondary degree based on survey data adjusted for non-response by the end of the study for the entire sample. This finding was found by both the

ED-PPSS analyses and the Mathematica analyses, but in Mathematica's case it was kept buried in an appendix to the Mathematica report and ignored in the text discussion in the disseminated Mathematica final report conclusions. The often-quoted Mathematica conclusion (Sefter et.al. 2009; Haskins and Rouse 2013, Decker 2013) *that UB had no discernible impact on postsecondary entrance or degrees earned is clearly incorrect.* The ED-PPSS analyses also found that with a balanced treatment and control group, the UB participants were 3.3 times as likely to obtain a bachelor's degree in six years compared with study participants who received no supplemental college access services (Cahalan and Goodwin 2014).

14. ***Don't Underestimate the Influence of Political Appointees in the Review Process and in Deciding Which Reports Get Published.*** The final Mathematica report was published by ED in the last week of the Bush Administration (January 2009) only upon direct orders from the departing political appointee staff. It was published over a year after the final contract formally ended in late 2007, and over the clear written formal objections of ED-Policy and Programming Studies Services (PPSS) career technical monitoring staff. It was also published after a formal "disapproval to publish" rating in the final Executive Secretary's (Ex. Sec) review from the Office of Postsecondary Education (OPE), out of whose appropriation the $14 million dollar study was funded. Its actual formal classification in the ED review process at the time it was published was "Returned to ED-PPSS for re-write." ED's formal publication review procedures were circumvented, and ED-PPSS was ordered in early January 2009 to get the report out by the end of the Bush Administration. While the ED-PPSS Technical Monitoring staff was not given details of the final negotiations with Mathematica, which were conducted in secret by the departing political staff, the reports were published with reported acquiescence and facilitation of the Institute for Education Sciences (IES). At the time, IES had a former Mathematica Vice President in a leadership position. In fact, in defiance of Mathematica's formal contractual agreements with ED, Mathematica had already published the Mathematica draft UB Evaluation final report on its own website in December 2008. After the Mathematica report was officially published by ED in 2009, PPSS attempted to get the results of their Quality Assurance Review also published by ED; however the publication of the QA results was blocked. Subsequently, ED-PPSS leadership gave permission to have COE publish the PPSS QA results in late 2009 (Cahalan 2009).

    In a further "suppression of transparency" after both Dr. Goodwin (the division director of the unit responsible for the study) and I had left the Department of Education, the UB study files were ordered withdrawn from availability for restricted release to other researchers. The files had been stripped of identifiers under a separate contract by RTI and made available under restricted release to other researchers in early 2011. The order to withdraw the data files came from the same former Mathematica senior staff person who had by this time become in charge of PPSS. This withdrawal of the data files was ordered in 2012, after having been released to only two sets of external researchers (University of Wisconsin and COE's Pell Institute). The researchers from the University of Wisconsin subsequently replicated the internal ED staff quality assurance results, also finding positive results for the Upward Bound programs.

The history of the Upward Bound evaluation shows how difficult it is for official Technical Monitors to do anything but rubber stamp findings by a respected, so-called "objective independent contractor." The Bush Administration's political appointees were looking for evidence that the federal TRIO programs were ineffective. In fact, the Assistant Secretary's representative to whom PPSS reported openly joked in regularly scheduled meetings with our PPSS group that my supervisors should stop me from working on the UB project. After raising concerns about the Upward Bound study, I experienced the classic responses to a "whistleblower" within the ED bureaucracy. As noted, after a review of the UB data, Dr. Goodwin, the original UB study Technical Monitor in the 1990s and then the ED-PPSS Division Director, also wrote a detailed memo to Mathematica leadership expressing his view that the UB study was "seriously flawed." However, he too was ignored by Mathematica, IES, and the ED political leadership.

15. ***In the Current Climate of Government, if a Study Can Be Used to Cut Services, It Will Be Used.*** Among some political operatives and researchers, it is the working hypothesis that government programs of the civil rights and War on Poverty era do not work, do not target the right persons, or at best are not worth the resources put into them. In 2011, Grover T. Whitehurst, former IES director, testified to Congress that Upward Bound and Head Start and similar programs had not been shown to be effective. The often-quoted Haskins and Rouse (2013) Brookings policy brief based on the Mathematica findings generalized to calling most existing college access programs "ineffective" and called for radical restructuring of the service programs into research demonstration evaluation projects. In November 2013, Paul Decker, President of Mathematica, went so far in his APPAM Presidential speech as to call those he labeled the "Youth Advocacy Community" as being guilty of "misdemeanors" and "felonies" for their legitimate expressions of concern relative to the proposed design for a new Upward Bound evaluation and for their successful lobbying in Congress to require that future federal evaluation studies in TRIO meet IRB requirements. More recently again in 2017, Paul Decker arrogantly used the Upward Bound results to argue that college access programs have not shown "clear evidence of effectiveness." In a circular process, he shamelessly cites the Haskins and Rouse proposals, which were justified by using the erroneous Mathematica flawed UB results, to argue for more research of the type conducted by his company on how to make the so-called "ineffectual" programs more effective.

16. ***Correct Mistakes, Do Not Cover Them Up.*** Several people have asked me why Mathematica, a well-respected firm—holder of the What Works Clearinghouse (WWC) contract--would be so unwilling to admit mistakes. I believe that part of the issue is that in the very competitive contract acquisition system it is very easy to overpromise and also to ignore basic questions about the validity and robustness of the proposed design to address the complex questions posed by the government. Once contractors have published high-profile study results, the perceived need to protect their own organizational reputation can make it very difficult to admit, even to themselves, that they may have made mistakes. In this case, practitioners and other stakeholders and ED-PPSS technical monitoring staff who raised questions were portrayed by Mathematica as being advocates who were anti-data, anti-random assignment, and ignorant of research methods.

Hopefully the "Upward Bound evaluation story" is an "outlier" in the history of evaluation research. It was a "perfect storm" example of a complex, long-running evaluation that "got it wrong" combined with political actors bent on zero funding the program and who welcomed any findings that would support the budget policies they were advocating. The Mathematica contractor, who also was the WWC contractor, refused to consider that they might have made mistakes. The contractor, with a strong reputation, also had powerful former staff allies in leadership positions in IES and later in PPSS. Given the current Trump Administration's proposals to cut program services, we ask the question, will there be more cases such as the Upward Bound evaluation? How can we avoid the mistakes of the Upward Bound evaluation? How can we foster feasible, ethical, transparent, accurate, and useful evaluations that are not harmful to those we are serving? Who is looking out for the rights of the stakeholders to the evaluations, especially in an era in which social service programs are targets for budget cutting or elimination?

REFERENCES

Cahalan, M. (2009). *Addressing study error in the random assignment National Evaluation of Upward Bound: Do the conclusions change?* Retrieve from http://www.pellinstitute.org/publications-Do_the_Conclusions_ Change_2009.shtml

Cahalan, M., & Goodwin, D. (2014). *Setting the record straight: Strong positive impacts found from the National Evaluation of Upward Bound, re-analysis documents significant positive impacts masked by errors in flawed contractor reports.* The Pell Institute for the Study of Opportunity in Higher Education, The Council for Opportunity in Education, June 2014. Retrieve from http://www.pellinstitute.org/publications-Setting_the_Record_Straight_June_2014.shtml

COE Request for Correction. (2012). Submitted to the U.S. Department of Education. Retrieved form http://www.coenet.us/files/pubs_reportsCOE_Request_for_Correction_011712.pdf

Decker, P. (2013). *False choices, policy framing, and the promise of "Big Data."* APPAM Presidential Address. Retrieved from https://www.mathematica-mpr.com/video/appam-presidential-address

Decker, P. (2017). *More than an axe: Use evidence to improve programs and policy*. Blog post on Mathematica web-site. Retrieved from https://www.mathematica-mpr.com/commentary/more-than-an-axe-use-evidence-to-improve-programs-and-policy?utm_source=SilverpopMailing&utm_medium=email&utm_campaign=New%20and%20Noteworthy%2005%2017%2017%20Revised%20(1)&utm_content=&spMailingID=17252104&spUserID=MTYwMjMyMTYwNgS2&spJobID=1021430613&spReportId=MTAyMTQzMDYxMwS2

Harris, D., Nathan, A., & Marksteiner. R. (2014). *The Upward Bound College Access Program 50 years later — Evidence from a national randomized trial.* University of Wisconsin-Madison Institute for Research on Poverty Discussion Paper No. 1426-14 (2014).

Haskins, R., & Rouse, C. (2013). *Time for change: A new Federal strategy to prepare disadvantaged students for college.* Brookings 2013.

Heckman, J., Hohmann, N., Smith, J., & Khoo, M. (2000). Substitution and dropout bias in social experiments: A study of an influential social experiment. *The Quarterly Journal of Economics*, May 2000.

Horn, L. J., Chen, X., & MPR Associates. (1998). *Toward resiliency: At-Risk students who make it to college.* U.S. Department of Education, Office of Educational Research and Improvement. Washington, DC: U.S. Government Printing Office.

IES, National Center for Education Statistics (NCES) Statistical Standards--- These may be accessed at the following site URL: http://nces.ed.gov/statprog/ IES, Joint Committee on Standards for Educational Evaluation (JCSEE) (2011). Program Evaluation standards (3rd. edition http://www.jcsee.org/program-evaluation-standards

Myers, D., & Schirm, A. (1996). *The short-term impacts of Upward Bound: An interim report.* Washington, DC: U.S. Department of Education, Planning and Evaluation Service.

Myers, D., Olsen, R., Seftor, N., Young, J., & Tuttle, C. (2004). *The impacts of regular Upward Bound: Results from the third follow-up data collection.* Report submitted to the U.S. Department of Education. Washington, DC: Mathematica Policy Research, Inc.

Myers, D., & Schirm, A. (1999). *The impacts of Upward Bound: Final report on Phase I of the National Evaluation.* Report submitted to the U.S. Department of Education. Washington, DC: Mathematica Policy Research, Inc.

Nathan, A.B. (2013). *Does Upward Bound have an effect on student educational outcomes? A reanalysis of the horizons Randomized Controlled Trial study.* A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy (Educational Leadership and Policy Analysis) at the University of Wisconsin- Madison 2013 Date of final oral examination: 02/08/2013

Seftor, N.S., Arif, M., & Schirm, A. (2009). *The impacts of regular Upward Bound on postsecondary outcomes 7-9 years after scheduled high school graduation.* Report submitted to the U.S. Department of Education. Washington, DC: Mathematica Policy Research, Inc.

Seastrom, M. (2002). *NCES Statistical Standards* (NCES 2003–601). U.S. Department of Education, National Center for Education Statistics. Washington, DC: U.S. Government Printing Office.

What Works Clearinghouse Standards. (2014). http://ies.ed.gov/ncee/wwc/pdf/wwc_version1_standards.pdfhttp://ies.ed.gov/ncee/wwc/references/idocviewer/doc.aspx?docid=19&tocid=1/

Whitehurst, G.T. (2011). *Testimony to 2005 Congress in IES Hearings*. November 2011

About the Author:

Dr. Margaret Cahalan is the Vice President for Research and Director of the Pell Institute for the Study of Opportunity in Higher Education, of the Council for Opportunity in Education (COE). Over a 30 year career she has directed numerous large sample surveys and evaluation studies. After working at Westat, Mathematica Policy Research and RTI, she joined the U.S. Department of Education from 2004 to 2011. In this role she served as the Leader for the Secondary, Postsecondary and Cross Cutting Division of the Policy and Program Studies Services (PPSS) that was responsible for the final contract of the Mathematica Upward Bound evaluation.

Contact Information: MARGARET CAHALAN, DIRECTOR, VICE PRESIDENT FOR RESEARCH, The Pell Institute for the Study of Opportunity in Higher Education, margaret.cahalan@pellinstitute.org