

Section II

Introduction: Journey Through Educational Research

Educational researchers would prefer to think that their trade is a precise, scientific discipline with well-defined concepts and standardized procedures leading to uncontested results. However, between the ideal and reality there is usually a wide gap. Social phenomena are generally too complex to be isolated and measured, rigorous research methods may clash with ethical concerns, and the search for objectivity may be clouded by program advocacy. Good researchers strive for a balance between what should be done (the “perfect” research) and what can be done. For those dealing with secondary sources, that is, research done by others, the negotiation between ideal and reality is even more frustrating. Jargon-laden research must be decoded into intelligible language, large amounts of work must be reviewed to select a few evaluations for inclusion, and at the end, the questions that propelled the search may remain unanswered.

The making of this report reflects all these challenges. The journey that started 18 months ago required reviews of hundreds of articles, reports, books, unpublished manuscripts, and other documents to produce the summaries included in this report. This chapter briefly describes the path traveled, its obstacles and discoveries along the way. (For a description of the report methodology, see *Overview and Research Note*)

The Journey and Its Obstacles:

The U.S. is perpetually awash in ‘new’ and self-proclaimed ‘highly effective’ programs for improving students’ academic achievement . . . The evidence that most of these programs ‘work’ has always been modest, and evidence of generalizability of effects is, for the majority of programs, non-existent (Sam Stringfield¹).

Finding evaluations of any quality is a difficult task, except for federal initiatives or grantee programs that mandate such studies. Program evaluation is a time-consuming process that may take money away from direct services. For many educators and youth program practitioners, already struggling with funding shortages, the idea that some of this money will be diverted from services to support research is anathema. However, without research, program practitioners may be perpetuating failing or mediocre interventions whose long-term consequences are much costlier to the young people and society. Although common sense indicates that interventions without a proven record of success should not be replicated, the search for the “magic solution” seems to overcome common sense. A non-scientific estimate of the literature search suggests a ratio of five “how to” reports – that is, reports on how to implement a specific but often untested intervention – to one evaluation of a program or strategy.

The search process for this report was particularly challenging, more so than for the two previous AYPF compendia. Over 200 documents were collected for an initial selection of less than 50. As described in the *Research Note*, the acceptance of evaluations for the report was dependent on five criteria that included population, measurements, methodology, length of research and scope. The following paragraphs discuss some of the obstacles encountered in satisfying the criteria and how they influenced this report’s outcome. For readers who are interested in research but not familiar with its basic terminology and standards, a brief explanation is provided at the end of the chapter, under the title “*Basic Principles of Educational Research.*” Definitions of research terms are included in the *Glossary*.

Population

The most important caveat about the documents reviewed was their treatment of the population. First, although the initial purpose was to include evaluations and programs for youth from all minority racial/ethnic groups, the final report includes few studies related to Native American and Asian/Pacific Island youth. The report's primary emphasis on African American and Latino youth reflects rather a lack of information on the other groups than a search process that focused on these two groups.

Second, most evaluations report on the student population as a homogeneous group, where demographics, such as race/ethnicity, appear as part of a description, but are rarely taken into account in the analysis. Few evaluations disaggregated their findings by sub-groups – ethnicity, gender, socio-economic status, English proficiency or baseline academic achievement. Disaggregating data requires more work during the data collection phase, demands a larger pool of students to provide statistically meaningful results, and risks exposing program weaknesses. However, this type of analysis is essential to highlight areas that require improvement and areas of proven success, thus offering key information for school administrators and program implementers.

The evaluations of *Chapel Hill-Carrboro City Schools* (CHCCS) and *GE Fund College Bound* are good examples of the value of disaggregating data. Results from the CHCCS program showed improved levels of proficiency in mathematics and reading for African American students with a reduction in the score gap between these students and their white peers. Yet the writing scores for African American students actually declined during the period of the study. The evaluation for the *GE Fund* indicated an overall increase in college enrollment rates for all participants, but more so for Latinos. However, the gap in enrollment between African Americans and whites increased. Faculty involved with the two projects can use the data to examine their strategies toward each group of students, to replicate the strategies that are boosting minority achievement and revise the strategies that

are not working. Programs that claim success without disaggregating their data may be helping one group of students while the other groups continue to fail. In fairness to the student population as a whole, these programs are not achieving their objectives.

Outcome Measurements

The initial criteria for acceptance of evaluations required a set of outcome measures that would provide a broad picture of the students' performance, such as test scores, number and type of credits taken, GPA, dropout and attendance rates, as well as postsecondary education or employment data. This requirement was based on the principle that relying on a single measure to assess a program may lead to incomplete, and many times, misguided conclusions. For instance, the evaluation of *Equity 2000*, a program that proposes academically challenging curricula for all high school students, shows a 30% increase in student enrollment in advanced mathematics classes. It also shows an increase of about 50% in failure rates in these same classes. While the enrollment data suggest an accomplishment, the data on passing rates indicate the need for much work before the program claims success.

Despite efforts during the search period, few evaluations reported more than two measures of achievement, the most frequent being test scores. It is inadvisable to use tests as the sole measure of student knowledge for many reasons. For instance, multiple-choice tests measure only one type of learning (memorization); some tests have been criticized as being culturally biased against minority students; some students are great test-takers while others are not; tests evaluate the student on one day out of 180 or more per school year, and on one set of specific competencies; tests do not necessarily assess the students' mastery of essential skills, such as problem solving, communicating complex ideas, using different strategies to reach a solution, or working in groups.² Notwithstanding the myriad problems with testing, the reality is that tests are being used across the country as a measure of school accountability and student achievement, and as gateways to advancement along the educational ladder.

Programs that raise the test scores of minority youth do increase the youths' chances of high school graduation, college admission and success in later life.

Acceptance of test scores as a valid measure of student achievement does not solve the question of whether it should be the **sole** measure. There are many methodological limitations associated with an overemphasis on test scores, such as:

- ◆ *Habituation* – Although questions change with different administrations of a test, students get used to the logic behind the test and its style. With time, scores in that specific test tend to go up due to habit, rather than actual improvement of student performance.
- ◆ *Lack of reliability* – Few of the current statewide tests are submitted to statistical analyses to assess their validity and reliability.³
- ◆ *Political pressure* – Tests may be weakened to address parental opposition and, in this case, increased test scores within a period of time may reflect a change in the tests (becoming easier or lowering the cut-off scores) rather than better-prepared students.
- ◆ *Teaching to the test* – Higher test scores may reflect the schools' emphasis on teaching to the test. With teachers focused on preparing the students to take the test, it is expected that scores will go up, even if the students still miss important competencies for future careers, sacrifice depth for breadth, and do not work on problem solving and critical thinking skills important for democratic citizenship and the new job market.
- ◆ *"Cheating" the system* – Higher test scores may also hide an increase in dropout rates or in the number of students identified as having limited English proficiency or in need of special education (generally, these students are exempted from statewide tests). As the students who test poorly for various reasons are pushed out of the system, the average scores of the remaining students increase. Without other

measurements, such as trend data on special education enrollment, dropout rate, college attendance and retention, enrollment in remedial courses in college, or type of employment, conclusions based solely on test scores are limited.

All this being said, with the current emphasis on testing it is understandable that researchers rely on tests to evaluate the success of a program. Indeed, the vast majority of evaluations found used test scores as the sole measure (at least, the sole quantifiable measure) to assess a program's performance.

Evaluations that use scores on only one test to assess a program create a serious obstacle for comparisons across programs. For example, is a 30% increase on the Texas Assessment of Academic Skills (TAAS) a greater feat than a 10% increase on the California Achievement Test (CAT)? Evaluations using the National Assessment of Educational Progress (NAEP) can be compared, since this is a nationwide assessment (although the NAEP is not conducted yearly and scores are not reported for individual students), but few studies reviewed used NAEP data.⁴ Another question that remains unanswered by a raw test score is its impact on the student's life. What does a three-point increase in a test represent for the student? Is this student now at the expected grade level? How much more does the student need to be proficient in the subject?

Translating results into grade levels or percentiles facilitates comparisons. For instance, after the *Calvert* model was implemented at the Dr. Carter Goodwin Woodson Elementary School, an all African American inner city school in Baltimore, the first grade average reading comprehension scores went up 31 points, from the 18th to the 49th percentile. This measure indicates that before *Calvert* was implemented, Woodson students were scoring on average below 82% of all students who took the Maryland test. One year into the new program, the average score of Woodson students placed them close to the middle. This information does not answer the question of how well the test

assessed what students need to know to succeed in life, but very few tests, if any, have such predictive power.⁵

Methodology

Design - Methodological rigor should be a concern for any researcher, but the standards of rigorous research are not so clear in the educational field. Evaluations using control or comparison groups were rarely found in the search for this report. The majority of the documents that we found compared the program or school with existing databases at the district or state levels. As the methodological rigor weakens, the findings become less reliable or generalizable, and the research process turns into an expensive, but fruitless exercise. Researchers who deal with limited budgets must carefully choose a design that provides the required information without unjustifiable expense. Interestingly enough, despite complaints of lack of research funding, the search produced a number of evaluations with highly complex, costly but inefficient designs.

Use of indicators – To evaluate performance changes, the data collected must be compared against either a baseline performance (how the students performed before the program) or some established indicator (how the students were expected to perform). A claim that 70% of Latino students in a program graduated is meaningless without information on how many students graduated before the program, or the overall graduation rate for Latino students in that specific school district or state. An enrollment of 80% in an algebra class may seem high until we discover that algebra is a mandatory course for graduation in that school district, and the enrollment should be 100%. Numbers gain meaning only within a context. This comment should be obvious, but a number of rejected evaluations claimed the success of a program without that context.

Statistical treatment of data – In addition to including baseline data and/or contextual indicators, researchers should calculate the statistical significance of their findings. A 12% decline in the test score gap between African Americans and white students in a specific program could reflect either

the positive impact of the program or normal fluctuations in test scores. Statistical tests are needed to separate random occurrences from treatment effects. If these test scores are performed, researchers must report results, including levels of significance. Again, reporting statistical significance is a basic research principle that was frequently forgotten among the documents reviewed.

Researcher bias – It is not uncommon in the educational field that a research institution or an individual researcher monopolizes the evaluation of specific programs or initiatives. In an ideal world, third parties (“outsiders”) with no direct interest in the program should conduct the evaluation to ensure the impartiality of analysis. In reality, however, it is often cheaper and easier for an “insider” or advocate with the appropriate research skills to conduct an internal evaluation. Fortunately, the review conducted for this report shows that “insider” evaluations can be just as rigorous and impartial as third-party evaluations. For example, many school evaluations are conducted through school district staff. Depending on the local political climate, these studies can be quite independent, particularly when they are intended as internal tools of assessment. The *Chapel Hill-Carrboro City Schools* evaluation is an example of an impartial insider research. In contrast, a number of “outsider” evaluations were rejected because they contained blatantly biased analysis.

Scope

If we do not describe the possible dystopias we shall be left only with [our] utopias. If we do not insist on bringing research findings (which may be politically inconvenient) into the public arena, we contribute to the erosion of democracy (Gipps⁶).

It is well-known that academic journals in any science (not only education) tend to publish evaluations that show success, while studies with negative findings are politely rejected. To ensure a more balanced perspective of programs geared toward minorities, the search included manuscripts

and unpublished grant reports in addition to published articles. Yet, whether the evaluation was published or not made little difference. A tendency to spin results into success or hide less than successful results was common to the majority of the documents. *Chapel Hill- Carrboro City Schools, GE Fund College Bound and High Schools That Work* deserve commendation for the courage to show accomplishments *and* shortcomings. Without this courage, program evaluation becomes little more than statistical cheerleading. Evaluators who hide negative results or use their trade as a tool for ideological positions are doing a disservice to policymakers, who will make decisions based on questionable information. By perpetuating misinformation, these evaluators are also doing a disservice to the educational process and to the youth, victims of failed strategies disguised as success.

The first conclusion resulting from the search process is that the most useful research is based on simple but methodologically sound design and provides information that is clear and easy to understand. This type of information is essential for educators and program practitioners who need to convince skeptics, placate critics, or expand support for their programs. Less useful are methodologically unsound evaluations, or evaluations that are so complex and hard to read that, high quality or not, they provide little usable information to policymakers and practitioners.

Report Overview

A brief overview of the evaluations selected for this report reflects the following characteristics:

- ♦ *Range.* The selected evaluations present a mix of policy initiatives and public or private programs. Together, the summaries span the educational ladder, from early childhood to graduate education. Although some district-wide reforms address all grades from K-12, evaluations of programs or initiatives that specifically target middle school students were not found. The search, albeit quite
- ♦ *extensive,* may have missed such programs, but this finding is worrisome, since many students who drop out of school start falling behind in middle school.
- ♦ *Population.* Few programs and initiatives target specific racial/ethnic groups. The majority serve a large number of minority students for two basic reasons. First, the majority of evaluations dealt with programs and initiatives targeting Title I schools, that is, schools with large numbers of students living at or below the poverty level. Although poverty is by no means an exclusive problem of minorities, minority children and youth are over-represented among the poor. Second, some programs are located in areas where a specific minority group predominates, such as Latinos in Puerto Rico and some schools districts in California, and African Americans in Washington, DC, and Baltimore. The Population textbox in each of the evaluation summaries in Section II reports the population in each study by racial/ethnic group, income level, geographical location, and program targeted level.
- ♦ *Methodology.* The studies summarized in this compendium vary in design and methodological rigor. Nineteen out of the 38 summaries use a control or comparison group, four are longitudinal studies, nine employ the pre/post-treatment method and eleven compare their findings against district, state or national databases (some use more than one method). Four summaries are descriptive only.
- ♦ *Measures.* For K-12 programs, test scores are the most common measure of academic achievement. Most evaluations rely on one type of test, often the state-mandated test. A few studies use standardized tests adopted nationwide, such as the Iowa Test of Basic Skills (ITBS) and the Stanford-9 (SAT-9).⁷ Among other indicators, high school programs frequently cite college enrollment data, while postsecondary education programs look at retention rates. Few

reports provide data on employment, including *Tribal Colleges*, *Compact for Diversity* and the three long-term studies of early childhood programs (*Abecedarian*, *Child Parent Centers* and *High/Scope*).⁸

This analysis discussed utopias and dystopias, the politically inconvenient but statistically significant. The hope is to contribute information that can guide educators and policymakers to better informed choices of strategies and initiatives that improve the academic achievement of minority youth; and foster a better understanding of the need for evaluation studies that look at facts, rather than dreams, and reality, rather than rhetoric. This hope is translated in the recommendations below.

Recommendations

- ♦ ***A large-scale, national and comprehensive educational research agenda must be developed*** to (a) determine which strategies and policies have resulted in the most benefit, for whom, and at what cost, (b) provide guidance to evaluators on what type of research would be most useful to policymakers and practitioners,

and (c) provide guidance to practitioners on why quality research is needed, how to initiate it and use it.

- ♦ ***Public and private funding sources must require and support high quality program evaluations*** and utilize findings to improve policy and practice, rather than to punish programs.
- ♦ ***Data must be disaggregated by race, ethnicity, limited English proficiency, disability status, gender and poverty level and be made publicly accessible*** to researchers, educators, policymakers, families and the public at large.
- ♦ ***Researchers should look into a range of achievement indicators*** including, numbers of students enrolled and dropping out, attendance, test scores, GPAs, graduation, suspensions, expulsions, and special education referrals. They should also translate their findings into language that is accessible to policymakers, practitioners, educators, families and students, so that research findings can be translated into better education policies and practices.

Addendum: Basic Principles of Educational Research

The next paragraphs attempt to provide readers who are not familiar with research with some very basic tools to help them navigate the summaries and use the findings to make their own assessment about the programs. These paragraphs reflect the many discussions about research among the members of the editorial team. However, its inclusion is not without a certain hesitation since a large amount of information is necessarily omitted.

Control Groups

The use of *control groups* provides the most rigorous design to assess the effect of an intervention, but it also raises important ethical questions. In educational research that uses control groups, two groups of individuals are randomly selected – one group attends the program (treatment

group) and the other does not (control group). When using a control group, the researcher ensures that the two groups are as similar as possible and limits the factors that may interfere with the education process. This control enables the researcher to attribute later differences between the treatment and the control groups to the program's effect with some degree of certainty (total certainty is an unattainable ideal). However, a control group supposes that the evaluators, with the consent of program directors or implementers, made a choice to provide a strategy that may help a group of needy youth while refusing it to another needy group, a difficult decision for any concerned individual. Programs that have more applicants than openings and select students through a lottery process have a natural control group in the students who do not win the lottery. The lottery is a totally random process that excludes the possibility of personal bias from admission personnel, but few programs use this system.

Comparison Groups

Evaluators can solve this problem in part by finding a *comparison group*, that is, an existing group of students similar to the treatment group who will not attend the program. For instance, students in two schools that are demographically and academically similar where one school implements the program and the other does not. A popular comparison in educational research is between students in a specific program and district wide, statewide or nationwide data. This type of comparison group is the easiest to identify, because the data already exists, but is the least reliable, since large databases include schools with different academic achievement, socio-economic background, type of personnel, and funding levels.

Matching

Control and comparison groups must be *matched* for demographics, socio-economic status, and prior academic performance to ensure that they are similar. If the groups are not matched on all these factors, the evaluators cannot infer whether the findings reflect program effect or the initial differences between the groups. A treatment group starting at a higher academic level than the comparison or control group is more likely to show higher scores even without the program. Or the converse may be true. The treatment group may have more students who are struggling academically. In this case, results may favor the control or comparison group even if the program is working. Although this explanation appears obvious, we found evaluation studies that claimed program success based on comparisons of groups that differ in their basic demographics and performance characteristics.

Pre- and Post-Treatment Data

Research using *pre- and post-treatment data* does not have the problem of group differences, but brings up other concerns, such as differences in tests used to measure progress, natural student maturation, or interferences due to the exit and entrance of students, changes in school personnel, and other factors.

Timing and Longitudinal Studies

Time is an important factor in evaluations. A study conducted too early, before the strategies are fully implemented, will not show clear results. Studies where the data is collected only once do not provide information about the program's ability to promote changes on an ongoing basis. It is not unusual that a program shows positive short-term changes as a result of the attention generated during its initial implementation. If this is the case, results may decline the following year, when the novelty has passed and the attention wanes. *Longitudinal studies* provide the best information to assess the program's performance. However, longitudinal studies are both difficult to implement and expensive. In addition, as the time passes, contact with research participants becomes more difficult, the initial treatment and control group dwindle, and results from such small samples become less prone to generalization. The *Abecedarian Project*, *Child Parent Centers* and *High/Scope Perry Preschool* are examples of the advantages and difficulties of long-term longitudinal research.

Use of Samples

In research, population is the generic name for what is being studied (it can be rats, as in experimental psychology research, as well as schools, students and teachers). Studies of small programs that exist in one school should include all the students as the results will be more reliable. However, for large studies, such as programs implemented in many schools or large school districts, it may become impossible to manage the study using the whole student population and the *use of samples* becomes imperative. In general, samples are randomly selected using some type of lottery, computer-generated numbers, or similar process. Researchers can also select samples to answer specific research questions. For instance, they can select only the best schools in a district to compare with the best schools in another district, or they can select only male students to analyze how a program affects males. When researchers select the sample, they should explain their selection process.

Sample Size

The size of the sample is important to ensure that results can be generalized to the total population. If the sample is too small, it may not be suitable for statistical tests. One of the problems with disaggregating data is that, when the total sample is divided, each sub-group must be large enough to provide statistically significant results. Terms such as large or small are relative to the initial size of the population and the type of study being conducted, including the questions asked and the type of tests required.⁹

Statistical Significance

After ensuring the quality of the comparisons, evaluators must also identify whether the results

have statistical significance, that is, where results cannot be attributed solely to chance. There must be some degree of confidence that the results can be attributed to the program. In educational research, a 95% confidence level is considered good; in medical research, where life and death are at stake, 5% uncertainty may be too much. This confidence statement can be expressed in levels of significance. A difference in test scores between two groups of students that is significant at the 5% level means that only 5 out of 100 students got that test score by chance. For the other 95, the change in grade is an effect of the program. Levels of significance (p) are generally written as a mathematical expression where $p \leq 0.05$ (for a 5% significance level) or $p \leq 0.02$ (2% significance level) and so on.

Following are 38 summaries of evaluations on programs and practices that influence the academic achievement of minority youth.

1. Stringfield, Sam. "Underlying the Chaos: Factors Explaining Exemplary U.S. Elementary Schools and the Case for High-reliability Organizations." In *Restructuring and Quality Issues for Tomorrow's Schools*, edited by T. Townsend. London: Routledge, 1993.
2. For a discussion of tests as measures of academic performance, see Bracey, Gerald. *Thinking About Tests and Testing: A Short Primer in Assessment Literacy*. Washington, D.C.: American Youth Policy Forum, 1999 (available at <http://www.aypf.org/BraceyRep.pdf>); Natriello, Gary and Aaron Pallas. *The Development and Impact of High Stakes Testing*. Paper presented at the High Stakes K-12 Testing Conference, sponsored by The Civil Rights Project, Harvard University, Teachers College, Columbia University, and Columbia Law School, 1998 (<http://www.law.harvard.edu/groups/>); Rotberg, Iris. "Five Myths about Test Score Comparisons," *School Administrator*, 53 (1996): 30-31, 34-35.
3. Validity refers to whether the test measures what it is supposed to measure (for instance, does the test measure the knowledge in English expected from a 5th grader in Texas?). Reliability refers to whether the test results can be replicated (do Texan 5th graders well-versed in English always score within a same range every time they take the test or are the results too unpredictable?). For more explanation on this topic, see Bracey, op. cit.
4. For a discussion of comparisons between TAAS and other tests, including the NAEP, see Jerald, Craig D. (2001). *Real Results, Remaining Challenges: The Story of Texas Education Reform*. Washington, D.C.: The Business Roundtable.
5. Bracey, op. cit., has a discussion on the use of the SAT on "predicting" student performance in college.
6. Gipps, Caroline. *The Role Of Educational Research In Policy Making In The U.K.* Paper presented at the American Educational Research Association (AERA) Conference, Atlanta, Georgia, 1993, p.16.
7. For explanations about the tests used in each evaluation, the reader is referred to the *Study Methodology* section at the end of each summary. For a brief description of the tests, please refer to *Glossary*.
8. Employment data in the Early Childhood evaluations was not included in the summary but can be accessed in the full document.
9. A very accessible, easy-to-read introduction to sampling is Sudman, Seymour. *Applied Sampling*. New York: Academic Press, 1976.

Evaluation Summaries

