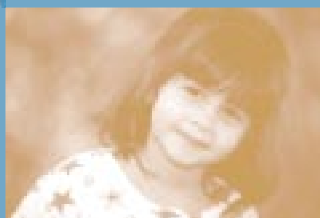


What it Takes:

Pre-K-12 Design Principles to Broaden Participation in Science, Technology, Engineering and Mathematics

April 2004

Board of Directors

Alfred Berkeley*
Former Vice-Chairman
NASDAQ

Kathleen Buechel
President and Treasurer
Alcoa Foundation

Mary Cirillo
Chairman and CEO
OPCENTER, LLP

Arnold Donald
Chairman and CEO
Merisant

Marye Anne Fox
University of California, San Diego

Augustine Gallego
Chancellor
San Diego Community College District

MRC Greenwood
Provost
University of California

Irwin Mark Jacobs*
Chairman and CEO
QUALCOMM

Anne Petersen
Senior Vice President for Programs
W.K. Kellogg Foundation

Deborah Wince-Smith
President
Council on Competitiveness

***Member of the Executive Committee**

National Leadership Council

Chair
Rep. Eddie Bernice Johnson (D-TX)

Bruce Alberts
President
National Academy of Sciences

George Boggs
President and CEO
American Association of
Community Colleges

Rita Colwell
Former Director
National Science Foundation

Malcolm Gillis
President
Rice University

Dain Hancock
President
Lockheed Martin Aeronautics
Company

Rep. Ruben Hinojosa (D-TX)
Chair, Education Committee
Hispanic Caucus

Erle Nye
CEO
TXU Corporation

Jill Sideman
President
Association for Women in Science

John Slaughter
President
National Action Council for Minorities
in Engineering

William Wulf
President
National Academy of Engineering

National Research Board

Eugene Garcia
Dean, School of Education
Arizona State University

Shirley Malcom
Head, Directorate for Education
American Academy for the
Advancement of Science

Willie Pearson, Jr.
Chair, School of History,
Technology and Society
Georgia Institute of Technology

Paula Rayman
Professor
University of Massachusetts

Richard Tapia
Professor
Rice University

Corporate Foundation Sponsors

Alcoa Foundation

American Institutes for Research

Bill and Melinda Gates Foundation

CH2M Hill

Hewlett Packard

Intel Foundation

W.K. Kellogg Foundation

Lockheed Martin

Microsoft

Pfizer

QUALCOMM

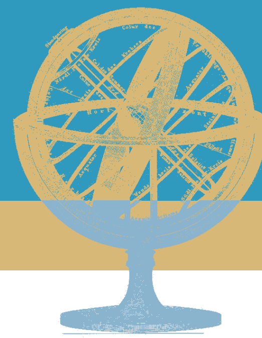
Richard Lounsbery Foundation

TXU

William and Flora Hewlett Foundation

The United States cannot afford to rely on a narrow and decreasing population segment to provide technical expertise that is the foundation of our nation's security, prosperity and quality of life.

BEST is a public-private partnership dedicated to building a stronger, more diverse U.S. workforce in science, engineering and technology by increasing the participation of underrepresented groups.



About BEST

BEST, an initiative of the Council on Competitiveness, was established as an independent, San Diego-based 501 (c)(3) organization in September 2001 at the recommendation of the Congressional Commission on the Advancement of Women in Science, Engineering and Technology. Our mission is to spur action to build a stronger, more diverse U.S. technical workforce. The nation's scientists, engineers, mathematicians and technologists comprise an indispensable strategic asset. Despite decades of effort, however, this pool of talent remains about three-quarters male and four-fifths white. The talent imperative we face is to prepare, attract and retain a larger share of all of our citizens in the technical fields that underpin U.S. economic strength, national security and quality of life.

BEST's objective has been to build a foundation for action through a two-year net assessment of best practices in pre-kindergarten through grade 12, higher education and the workplace to increase the participation of women, African Americans, Hispanics, Native Americans and persons with disabilities in the science, engineering and technology professions. Three blue-ribbon panels have worked in parallel across the whole continuum of education and workforce development with the guidance and support of BEST's Board of Directors, National Leadership Council, Research Board and Project Integrators who are listed on the inside front and back covers of this report.

Based on available research evidence and the professional judgment of 120 nationally recognized practitioners and researchers, the assessment:

- makes the case for national action to meet the U.S. talent imperative;
- rates pre-K-12 programs that have research evidence of effectiveness or are worthy of investment in further research;
- analyzes higher education and workplace exemplars;
- distills the design principles that underpin effective programs; and
- proposes an action agenda at the national and community levels to engage employers, educators, policy makers, professional societies and nonprofit organizations.

BEST will report its findings and recommendations to members of Congress in the spring of 2004.

BEST
Building Engineering & Science Talent

401 B Street, Suite 2200
San Diego, CA 92101
(619) 615-2940
www.bestworkforce.org

Table of Contents

Acknowledgments

This report reflects generous contributions of time and expertise from many committed individuals. In particular, BEST wishes to thank Dr. Shirley Malcom and Mr. Alfred Berkeley for their leadership as co-chairs of the blue-ribbon panel on Pre-K-12 education; Dr. Eugene Garcia for his expert guidance as advisory panel leader; Dr. Linda Rosen for her pivotal role in conceptualizing the issues through numerous iterations of the report; Dr. Elizabeth Cohen and Dr. Douglas Carnine for their major contributions to the Protocol developed by the panel to rate evidence of effectiveness; Drs. Carlos Rodríguez, Bruce Christenson and Marian Eaton of the American Institutes for Research for the rigorous development and application of the Pre-K-12 Research Protocol and presentation of the findings; Mr. Bruce Boston for his valuable assistance in framing the issues; and all of the project integrators and panel members who participated in the nominating and screening process and commented on successive drafts. Dr. Joanne R. Gribble thoughtfully edited the report, for which BEST alone assumes responsibility.

Executive Summary

The thinning of the U.S. technical talent pool starts early

This report focuses on what is working in pre-K-12 education to meet a difficult and increasingly urgent national challenge — the underrepresentation of women, African Americans, Hispanics, Native Americans and persons with disabilities in science, technology, engineering and mathematics. These underrepresented groups comprise nearly two-thirds of the overall U.S. workforce yet only make up one-quarter of the science and engineering workforce.

Too few young Americans do well in mathematics and science, according to international comparisons. Interest in many of the careers based on these disciplines also has declined. Moreover, the United States cannot afford to rely on a narrow and now decreasing segment of the population in technical fields that are the foundation of our nation's security, prosperity and quality of life.

The thinning of the U.S. technical talent pool starts early. Unequal access to educational opportunities, a pernicious achievement gap in reading and mathematics, differential rates of high school graduation and lack of encouragement are all key contributing factors. But there are a few bright spots. Some students from underrepresented groups do acquire the foundational skills and confidence to move forward in technical disciplines. In the best of worlds, success would be easily recognized, and readily expanded. In the real world, knowledge of what works is a scarce resource and effective programs do not necessarily grow or serve as models.

The Search for Evidence

To help prepare more students from all groups in mathematics and science, **Building Engineering and Science Talent** (BEST) assembled a blue-ribbon panel of experts with broad experience and varying perspectives. The panel set itself the goal of rating program effectiveness based on available research evidence.

Such evidence is hard to find. Many programs concentrate their limited resources, understandably, on providing services and recruiting participants rather than on rigorous and costly impact studies. Program features such as selective recruitment and enrichment outside the classroom further complicate the task of research. Some programs do not survive long enough to demonstrate their effectiveness.

The BEST panel, in collaboration with the **American Institutes for Research** (AIR) developed a protocol to guide an in-depth evaluation of research evidence. The protocol defined a rigorous study as one that provides meaningful research evidence comparing the outcomes of students who experience a given intervention and those who do not. Applying the collective professional judgment of the national panel of experts, an initial list of 200 programs were nominated for consideration. The 34 programs ultimately selected for rating represented a cross section of interventions varying in content, purpose, grade level, mode of delivery and sponsorship. Most were targeted to the needs of one or more underrepresented groups, but some were aimed at all students.

To be considered **verified**, the protocol requires five studies of acceptable rigor conducted by independent evaluators to have substantially positive results. The protocol classifies a program as **probable** if two or more acceptably rigorous studies by independent researchers show positive immediate or longer-term effects and there are no substantially negative studies. A program is considered **notable** if at least one acceptably rigorous study is substantially positive and no studies have substantially negative results. Programs having only descriptive evidence of success in serving one or more of the BEST target populations are rated as meriting **further research investment** on a priority basis. Finally, some programs currently **lack the necessary evidence** to be evaluated for rigor, and are not described here.

Findings

Probable Effectiveness

Twenty of the 34 interventions examined by AIR yielded research-based or descriptive evidence. Although none of the programs examined had enough evidence to be considered **verified**, two earned the rating of **probable**:

Direct Instruction (in Mathematics), an instructional approach developed in the late 1960s at the University of Illinois, provides intense, sequenced lessons whose aim is to allow all children — even the lowest performing — to master academic skills in mathematics. DI's mathematics programs are used in schools nationwide, many of which are low-performing schools in high poverty areas, including both urban and rural districts. Research evidence comparing DI and non-DI students indicate that DI has a positive immediate effect on mathematics performance on standardized tests, although whether that persists long-term is yet to be examined.

Project SEED (Special Elementary Education for the Disadvantaged) a supplementary mathematics program begun in the 1960s in Berkeley, California, targets urban students in grades 3-6. It aims to improve students' critical thinking and problem-solving skills, achievement levels in mathematics and academic self-confidence. Project SEED operates in several large urban school districts serving minority and disadvantaged students. Research evidence found that students who participated in the Project SEED for a semester outperformed their non-SEED counterparts on standardized mathematics tests. Students who received several semesters of SEED continued to outperform non-SEED students for at least four years and went on to take more higher-level mathematics classes in middle school and high school.

BEST identified a framework of design principles or shared features from programs with research-based evidence of effectiveness.

Notable Effectiveness

Seven programs met the criteria to be rated as notable, based on the results of at least one rigorous study as well as descriptive evidence of effectiveness:

Advancement Via Individual Determination (AVID) is an in-school elective class program established in 1980 in San Diego City Schools. It seeks to enable at-risk middle and high school students who are promising but underachieving to succeed in a college preparatory program, and ultimately to enroll in college.

The Algebra Project provides the conceptual understanding of mathematics necessary to complete algebra successfully and enter the high school math sequence required for college entrance. It targets inner city and rural students in the upper elementary and middle school grades, particularly in high-poverty, high-minority communities.

Foundational Approaches in Science Teaching (FAST), an inquiry-based science curriculum series for students in grades 6-10, provides a three-year sequence of courses focused on physical science, ecology and the interrelationships among science, technology and society.

Gateway to Higher Education, an outreach program of the State University of New York (SUNY), prepares New York City high school students for studies and careers in science, technology, engineering, medicine, health and the teaching of mathematics.

Project GRAD (Graduation Really Achieves Dreams) aims to reduce dropout rates and increase college enrollment of minority students.

Puente, a California-based college access program, provides a bridge for Latino students from high school to college and from community colleges to four-year colleges.

Yup'ik Mathematics, an elementary school program in southwestern Alaska, seeks to improve the thinking and performance in mathematics of Yup'ik students while reinforcing Yup'ik culture.

Descriptive Evidence Warrants Further Research

Eleven additional programs yielded descriptive evidence sufficiently robust to call for further research on a priority basis. These programs are:

- American Chemical Society (ACS) Project SEED
- Detroit Area Pre-College Engineering Program (DARCEP)
- Disabilities, Opportunities, Internetworking and Technology (Do-It)
- El Paso Collaborative for Academic Excellence (EPCAE)
- Family Mathematics
- MATHCOUNTS
- Mathematics, Engineering, Science Achievement (MESA)
- Operation SMART – Girls Inc.
- Texas Pre-freshman Engineering Program (TexPREP)
- University of North Carolina Mathematics and Science Education Network (MSEN)
- Xavier University Summer Science Academy

Implications for Policy and Practice

The baseline of knowledge regarding what works established in this report is only a starting point. From there, BEST's experts have been able to develop a list of implications for policy and practice with the potential to broaden the base of pre-K-12 math and science education nationwide.

1. Distilling usable insights. BEST identified a framework of design principles or shared features extracted from programs with research-based evidence of effectiveness. These design principles are not a causal explanation of what works, yet are reasonable inferences of what it takes to succeed:

- **Defined outcomes** drive the intervention and are successfully accomplished for the entire target population. Students and educational staff agree on goals and desired outcomes. Success is measured against the intended results. Outcome data provide both quantitative and qualitative information. Disaggregated outcomes provide a basis for research and continuous improvement.
- **Persistence** enables effective interventions to take hold, produce results, adapt to changing circumstances and persevere in the face of setbacks. The conditions that ensure persistence include proactive leadership, sufficient resources and support at the district and school levels.
- **Personalization** acknowledges the development of students as individuals as the goal of the intervention. Student-centered teaching and learning methods are core approaches. Mentoring, tutoring and peer interaction are integral parts of the learning environment. Individual differences, uniqueness and diversity are recognized and honored.
- **Challenging content** provides the foundation of knowledge and skills that students master. Curriculum is clearly defined and understood. Content goes beyond minimum competencies, relates to real world applications and career opportunities, and reflects local, state, and national standards. Students understand the link between content rigor and career opportunities. Appropriate academic remediation is readily available.
- **Engaged adults** who believe in the potential of all students provide support, stimulate interest and create expectations that are fundamental to the intervention. Educators play multiple roles as teachers, coaches, mentors, tutors and counselors. Teachers develop and maintain quality interactions with students and each other. Active family support is sought and established.

While these design principles do not account for the success of any particular intervention, they represent a checklist of the essential characteristics of programs that have passed through a research filter. Viewed as a whole, they yield a number of insights that have the potential to inform policy as well as practice:

- The design principles of programs that work appear to comprise a package rather than an à la carte menu.

- The fundamentals associated with programs that target students from underrepresented groups seem apt to work for all students.
- The design principles derived from math and science programs appear to be applicable to other disciplines.

2. Deepening the knowledge base. The fact that a substantial search for effective programs turned up relatively little research evidence is in itself an important finding. The lack of reliable evidence is by no means limited to mathematics and science education. The national interest lies in deepening the research base as quickly and cost effectively as possible.

Government, industry and philanthropies that fund mathematics and science programs targeting underrepresented groups should set an aggressive leadership example in this effort. While a firmer understanding of what works does not provide a magic bullet, it can guide and improve the allocation of resources, support more informed decisions by educators, students and their parents, and provide the foundation for vitally needed continuous improvement. Based on lessons learned from this report, a realistic, well-focused research agenda should include:

- An ongoing nationwide application of the BEST/AIR protocol to inventory and analyze the research on mathematics and science interventions that focus on groups underrepresented in these disciplines;
- Ongoing analysis of additional research and evaluation information on the programs identified to date.
- Creation of a resource pool to support this research agenda by establishing research evaluation partnerships encompassing congressional, federal, state, local and private sector interests.

3. Tightening the links between research, policy and practice. The objective of deepening the knowledge base is to inform the operational decisions of policy makers and practitioners by providing points of reference that are empirically grounded. Unless research findings are easily accessible and seen to matter, they are not likely to draw the attention of a harried superintendent or principal on a regular basis. The same considerations hold for many policy makers, particularly members of Congress. BEST recommends that congressional hearings on the interrelationship between policy, practice and research in pre-K-12 mathematics and science education focus on supply issues to define and fund a priority national agenda that meets the national demand by connecting the three domains more closely.

4. Aligning system-wide and targeted approaches. The interventions examined by AIR fall into two broad categories — those aimed at "lifting all boats" (system-wide) and those tailored specifically to one or more underrepresented groups (targeted approaches).

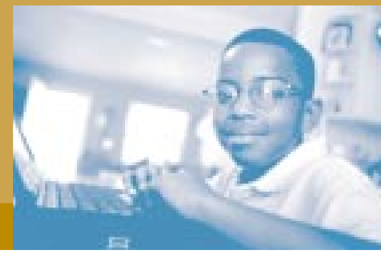
The underlying premise of "lift all boats" is that the best way to strengthen mathematics and science education for some students is to do so for all. Its *system-wide* focus is on leverage points that have broad applicability nationwide such as teacher preparation and professional development, curriculum, classroom environment, standards and accountability. *Targeted approaches* hold that the pre-K-12 system alone cannot meet the challenge. These interventions are required to meet the distinctive and widely varying needs of students from underrepresented groups by providing support, motivation and learning in specific contexts.

Whereas a system-wide focus represents the norm in school districts across the country, targeted programs are typically established at the margin of the pre-K-12 system with external funding — a combination of government, corporate and foundation grants known as "soft money" — and thus subject to the perils of discretionary funding practices.

These approaches must be more closely aligned. System-wide approaches have a limited capacity to meet challenges at the individual level, while targeted programs have a limited capacity to bring solutions to scale. The design principles extracted from programs with research-based evidence have great relevance system wide. A more comprehensive dialogue between program developers that are succeeding "in the trenches" and decision-makers with system-wide objectives and responsibilities could:

- develop an action plan to insure that programs that work do not remain at the margins of the system;
- increase the capacity of pre-K-12 mathematics and science education as a whole to develop the talent of students from all groups.

Chapter 1



Broadening the Base

Bridging the gap between "what's wrong" and "what it takes"

We are not getting it done. Although innovation is the lifeblood of the U.S. economy, we are not developing sufficient scientific, engineering and technical talent to meet our own needs. Instead, we have come to rely increasingly on imported brainpower in our research laboratories, product design centers, software development houses and universities. These transfusions of talent are no longer a supplement, but a necessary part of our innovation enterprise. Their value is beyond question, but our dependence reflects weaknesses and vulnerabilities that cannot be overlooked.

One part of the solution is getting all American youngsters off to a stronger start in mathematics and science. Viewed broadly, the math and science education now being delivered in the years between pre-kindergarten and 12th grade, like a sterile vaccine, is not "taking" in enough American classrooms. Our students are not learning enough of what they need to know to move into scientific and technical fields where this knowledge is foundational. The deficiencies in pre-K-12 math and science education have been amply documented in any number of nationally significant reports.[1]

The bridge between "what's wrong" and "what it takes" is still very much under construction. The most recent national assessment measuring the progress of American students over time points to significant long-term gains in mathematics among fourth and eighth graders. Samples in every state show broad improvement in the performance of all groups from 1990 to 2003. Still, 68 percent of fourth graders and 71 percent of eighth graders scored below the "proficient" level while no more than 5 percent achieved at the "advanced" level. The positive direction of change has not yet altered the realities of overall underachievement, nor has it left students of color on equal footing.[2]

The picture looks grim in international comparisons, too. Paradoxically, even though the United States is a world leader in science and technology, the results show that American students lose competitive ground vis-á-vis their international counterparts as they move through school. Among 20 nations assessed in advanced mathematics and physics in 1995, the United States ranked near the bottom of the pack.

This fact alone helps explain why increasing numbers of American students shy away from science and engineering. Aside from a positive trend in the life sciences, undergraduate and graduate degrees granted to American students in math, physics, chemistry, computer sciences and engineering remain below levels reached in the early 1990s. This drop-off has taken place even though there are more students enrolled in higher education than ever before. Meanwhile, growing numbers of well-trained international students have streamed into U.S. graduate schools of science and engineering, earning 40 percent of the Ph.D.s in 2002. Industry and research universities have drawn heavily on this talent pool. The share of foreign-born science and engineering workers in the U.S. technical workforce increased from 11.2 percent in 1980 to 19.3 percent in 2000, reaching a total of 1.5 million.[3]

Who will replace the nation's current science and engineering workforce, one-quarter of which is more than 50 years old? What will become of those high-paying jobs when there is not a sufficient pool of homegrown talent to fill them? What does all of this mean for America's prosperity, security and quality of life? These are the ultimate stakes of pre-K-12 math and science education.

With a mandate from the U.S. Congress, Building Engineering and Science Talent (BEST) has spent the past year focusing on what it will take to tackle one large piece of America's problem in developing technical talent: how to increase the participation of groups that are traditionally underrepresented in science, technology, engineering or mathematics. These underrepresented groups – women, African Americans, Hispanics, Native Americans and persons with disabilities – comprise the majority of both the population and the workforce. Yet as Table 1-1 shows, they are not entering the technical fields that have relied historically on a narrow and decreasing segment of the population, white males.

If the trends of the past 25 years hold, girls in today's pre-K–12 classrooms will be the most highly educated segment in American society but they will stay away in droves from science and technology. African Americans, who also have made important gains in education and income since the mid-1960s, still represent just 3.5 percent of the science and engineering workforce. Hispanic

Table 1-1
Proportion of Employed Scientists and Engineers in the U.S. by Broad Occupation, Race/Ethnicity, Sex and Disabilities, 1999

Sex, Race/Ethnicity and Disabilities	Percentage of U.S. Population 1999	Total Occupations	All S&E Occupations
White men	35.2	39.9	63.2
White women	37.7	34.8	18.6
Asian men	1.8	2.0	8.4
Asian women	2.0	1.8	2.6
Black men	5.7	4.9	2.1
Black women	6.4	5.9	1.3
Hispanic men	5.8	5.9	2.4
Hispanic women	5.7	4.2	1.0
Native American men	0.4	N.A.	0.2
Native American women	0.4	N.A.	0.1
Persons with disabilities	2.0	0.7	5.8

Source: CPST, data derived from National Science Foundation
Note: Totals may not add to 100 due to rounding.

Americans, the largest and fastest-growing minority, account for 25 percent of the current school population and are similarly underrepresented in science and engineering careers. Native American youngsters are also being left on the sidelines, as are students with disabilities for whom science and technology often serve as a vital lifeline. These groups comprise America's underrepresented majority.

It is their education in pre-kindergarten through 12th grade that will play a critical role in meeting the challenge of underrepresentation. Most of the thinning of the pool of potential technical talent occurs by the end of high school. By then, students have acquired the foundational skills and confidence to pursue further study in technical disciplines or they have not. They have been encouraged and shown interest in science, math and engineering or turned off.

Clearly, our best chance to meet America's technical talent imperative is to train the promising young people we already have. More than half belong to these underrepresented demographic groups and, as we have seen, they underperform or opt out of math and science education in great numbers. This report summarizes what BEST has learned about what it takes to beat the odds and what it will take to improve them for all students.

Two Basic Approaches

BEST has found two basic approaches to what it takes to provide better math and science education to all students, including girls, students of color and the disabled.

Lift all boats The first approach has been to try to "lift all boats," as most recently urged in the Glenn Commission report[4] and the Bush Administration's "No Child Left Behind" legislation. System-wide strategies come in many shapes and sizes, but all share the fundamental belief that fixing math and science education is a national problem requiring a national effort. The underlying premise of "lift all boats" is that the best way to strengthen math and science education for some students is to do so for all. Proponents therefore seek to generate an impact that improves the entire education delivery system. From their standpoint, this is the surest, most cost-effective way to provide the foundational skills and motivation to equip the underrepresented majority for the next level of education or a skilled job.

Typically, the system-wide approach has focused on broad national remedies, e.g., improving teacher preparation, upgrading professional development, restructuring instructional environments, devising cutting-edge curricula, raising expectations for student performance, insisting on clear and rigorous standards and urging more (and more exacting) accountability for teachers, principals, administrators and policy makers.

Thoughtful advocates of "lifting all boats" acknowledge the inequities embedded in pre-K-12 education and the impracticality of one-size-fits-all solutions. Many recognize that pervasive disparities in funding, the lack of trained teachers and limited access to learning opportunities put low-income, mainly minority students at a huge disadvantage. Indeed, the achievement gap between underrepresented minorities and non-minority students in math and English is widely viewed as a defining challenge for the system.

Even so, the objective of making as broad an impact on as many students as possible is in its own way a limiting factor. It compels a search for leverage points that affect the entire system. The current national discourse and policy favors school performance outcomes as viable measures to drive educational improvement. Other ongoing national efforts address increasing the national supply of qualified math and science teachers as key; promulgating curriculum and student performance standards; the small-schools movement concentrating on school size. These leverage points, by definition, are means of pursuing multiple objectives. Improving math and science education for underrepresented groups is one priority, but not the goal that shapes overall strategy.

Targeting the underrepresented The second broad approach starts from the premise that the pre-K-12 system alone cannot meet a challenge as urgent and demanding as that of nurturing the talent of underrepresented groups in math and science. These fields are not popular for most American youngsters, especially for those who do not see ready role models, receive encouragement at home or have any idea what a scientist or engineer really does. Success requires a level of attention, support, motivation and real-world exposure that the system as a whole cannot be expected to provide. The system can help, but where it falls short, additional intervention is required – whether to give a girl in the suburbs the confidence to take chemistry, provide an after-school learning opportunity in a low-income school, or find a math mentor for a student who would benefit from an adult's concentrated attention.

The targeted approach aims to make a difference with specific populations whose circumstances and needs vary widely. African American, Hispanic and Native American youngsters face distinctive challenges that require carefully tailored strategies. Non-minority females who do not see science and technology as a possible career pose a different challenge than Latinas who may not be expected to go to college at all. The problems facing students with disabilities demand yet a different set of interventions. Targeting seeks to make an immediate impact through a delivery instrument that is programmatic rather than systemwide. The point is not to "raise up" all students but to get as many girls and boys, African Americans, Hispanics, Native Americans and students with disabilities on a trajectory of success in science- and math-based fields.

Typically, targeted programs are established at the margin of the pre-K-12 system. While some are delivered in classrooms, their principal funding is often external – a combination of government, corporate and foundation grants known as "soft money." While targeted programs have changed thousands of lives and have contributed in fundamental ways to the diversity of America's

Ultimately, the nation cannot afford to choose between systemic and targeted approaches.

current pool of technical talent, their limits also have to be recognized. The financial and leadership bases of many interventions are fragile. Collectively, these interventions lack the reach to deliver results on the scale that our country needs.

Ultimately, the nation cannot afford to choose between the systemic and targeted approaches. Clearly both are needed. Until now, however, the two have not been bridged effectively. Their proponents – both equally committed – seem worlds apart. If trying to move a system of 50 states and 15,000 school districts is a Herculean task, so also is building and sustaining a program at the periphery of the system that enables a female, a minority student, a person with a disability or a first-generation American to get into a top engineering school. But system-wide approaches have a limited capacity to meet challenges at the individual level, while targeted programs have a limited capacity to bring solutions to broader scale.

BEST believes that it is both imperative and possible to create a mutually reinforcing relationship between these approaches. In the best of worlds, efforts to lift all boats will be informed by a deeper understanding of what it takes to succeed at the level of the targeted program, while effective targeted programs will be more tightly integrated into the system.

Moving toward Convergence

This report lays groundwork for convergence of the two approaches in two ways. First, it tackles the fundamental question of evidence – *what evidence confirms that an intervention is effective in preparing and motivating students from underrepresented groups in math and science?* Everyone concerned with making headway on this issue — policymakers, educators, researchers, private sector funders, parents and students — must make decisions based on knowledge rather than on anecdote or intuition.

The blue ribbon panel of experts assembled by BEST, together with the American Institutes of Research, has developed and applied a first-of-its-kind protocol to measure the quantity and quality of evidence for targeted programs or systemic initiatives. In a complex field where building agreement over educational methodology is challenging, the panel reached consensus on how to determine with varying levels of certainty that a given intervention works. Chapter 2 briefly describes the substance and application of the protocol. The protocol itself appears as Appendix 1.

The resulting evaluations, drawn from a broad sample of nominated programs, identify nine programs currently in use in some education settings nationwide that have some level of research-based effectiveness, and 11 others with sufficient promise to warrant further research. Chapter 3 provides capsule summaries of the programs that meet these standards.

The second means by which advocates of the differing approaches can come together is through the design principles or shared features that BEST has extracted from programs with evidence of effectiveness. We believe the set of insights laid out in Chapter 4 has equal usefulness for both the targeted and systemic approaches. Yet they are by no means a causal explanation of what works. They are necessarily inferential, but are reasonable inferences. They provided an approximation grounded in knowledge and evidence of what it takes to succeed with underrepresented groups.

Chapter 5 presents a users' guide to the effective and promising programs that were identified during the BEST/AIR study process. We believe it holds promise as a means to increase the body of knowledge about helping underrepresented students in the math and science pipeline.

The dialogue on broadening the base for pre-K-12 math and science education must include and integrate the national and local levels. The pre-K-12 BEST initiative seeks to frame that dialogue.

ENDNOTES

1. Committee for Economic Development. (2003). *Learning for the future: Changing the culture of math and science education to ensure a competitive workforce.* New York: Author. [On-line]. Available : www.ced.org/docs/report/report_scientists.pdf; National Commission on Mathematics and Science Teaching for the 21st Century. (2000, September). *Before it's too late*. Washington, DC: U. S. Department of Education. [On-line]. Available: www.ed.gov/americacounts/glenn; National Council of Teachers of Mathematics (2000.) *Principles and standards for school mathematics*. Reston, VA: Author. [On-line.] Available: standards.nctm.org/document/index.htm; National Committee on Science Education Standards and Assessment. (1996). *National science education standards*. Washington, DC: National Academy Press. [On-line]. Available: www.nap.edu/catalog/4962.html; National Commission on Teaching & America's Future (1996). *What matters most: Teaching for America's future*. New York: Author. [On-line]. Available: www.teachingquality.org/resources/pdfs/What_Matters_Most.pdf

2. National Center for Education Statistics. (2003). *The nation's report card*. Washington, DC: U. S. Department of Education. [On-line]. Available: nces.ed.gov/nationsreportcard/

3. National Science Board. (2002, January). *Science and engineering indicators — 2002* (NSB-02-1). Arlington, VA: National Science Foundation.

4. National Commission on Mathematics and Science Teaching for the 21st Century.

Chapter 2

Weighing the Evidence

Evidence matters because outcomes matter and resources are limited

The Blue Ribbon Panel on Best Practices in Pre-K-12 Education was assembled by BEST to identify what works to broaden the base of successful students in pre-K-12 math and science education. The panel brought together practitioners, policy makers and researchers with wide-ranging experience and viewpoints. All members shared a commitment to increase the success rate of underrepresented groups in math, science and pre-engineering. All members recognized the importance of evidence in answering the question, "What works?"

While hundreds of elementary and secondary education programs have been implemented to advance, prepare and motivate students in math and science, judgments on their effectiveness are often anecdotal. To this end, BEST examined systematically the literature for the evidence of the effectiveness of specific interventions. This chapter describes this process and highlights the importance of "empirical" (i.e., systematically observed and measurable) evidence of program effectiveness while providing a protocol by which available evidence can be evaluated.

The interventions selected for review do not represent an exhaustive list. Rather, they represent those efforts known to BEST, the Blue Ribbon Panel on Pre-K-12 Education and the consultants at the American Institutes of Research (AIR), who were engaged to help develop an evaluation procedure. These embrace the science, mathematics, engineering and technology disciplines as well as a variety of grade levels for most of BEST's targeted groups. Some focus on particular subjects while others have a more general purpose of preparing students for college. Similarly, some are designed to target particular groups of students while others are thought to work well for a broad spectrum of students.

Why Does Evidence Matter?

Evidence matters because outcomes matter and resources are limited. It is important to ensure that investments in money, time and human capital have a high probability of paying off. Good intentions and passionate commitment are not enough to fill the science and technology pipeline, especially with groups of students which have been largely left out up to now.

At present, reliable empirical evidence is hard to find. Indeed, authors of previous analysis have commented on the lack of well-designed statistical and research-based evaluations that assess the effectiveness of programs. Several reasons may account for the paucity of a solid research base.

- First, it is difficult, costly and time consuming to conduct rigorous studies of a program's effects. Many programs appropriately concentrate their limited resources on providing services and recruiting participants rather than on impact studies. Research by third parties often requires sponsorship by a government agency or foundation.
- Second, the particular features of programs such as selective recruitment, informal participation and multiple interventions may make it difficult to design a well-controlled and defined study.
- Third, programs led by charismatic leaders, or tied to new federal, state or local initiatives, may not survive long enough to demonstrate their effectiveness.

Although most programs operating in more than a few schools can produce at least one enthusiastic endorsement, educators and policy makers need more than anecdotal evidence and the marketing materials of developers to validate effective programs. Thus, despite the challenges the empirical evidence for effectiveness is needed.

Development of Protocol to Evaluate Empirical Evidence

BEST's goal was to rate the available evidence of a program's effect on underrepresented students' outcomes and identify those that had a sufficient robust research basis. The process yielded three categories: those with an acceptable research basis, those that would benefit from rigorous evaluation and those for which sufficient evidence is not available. Those in the first category were further subdivided into three groups, verified, probable and notable, in order to reflect different levels of convincing evidence.

A description of the process and ratings follow, but a word about how the evaluation protocol was developed should be made at this point. Many viewpoints on education and industry were represented on the Blue Ribbon Panel on Pre-K-12 Education. It is thus not surprising that there were many different opinions on how the existing research should be evaluated. In the end, the panel engaged AIR to help develop a procedure for evaluating the available evidence on success in pre-K-12 science and math education that represented a consensus among the different perspectives. The objective was to provide a multi-category rating system that accomplished three goals: balance (1) the need for high standards of evidence with (2) recognition that evidence would probably take a variety of forms and (3) the continuing need to encourage improvements in program evaluation.

To achieve this goal, the review procedure — called by BEST the protocol — was built on elements of similar instruments developed for other research reviews, but was tailored to particular concerns voiced by members of the BEST panel:

- The protocol is limited to research evaluations that have a comparison group.

• The protocol focuses on questions central to maintaining a high standard for evaluating the quality of research studies (e.g. research design, sampling strategy, etc.).
• The protocol's evaluation of program effectiveness reflects intermediate and long-term outcomes as well as immediate effects.
• The protocol considers small studies since some of the interventions reviewed are small programs with focused efforts.
• The results of the protocol are explained in non-technical language in order to be accessible to policy makers and the public.

BEST panelists and staff reviewed and commented on a draft protocol developed by AIR. Following revisions to the draft protocol, a pilot test was conducted on available research for seven programs. Additional adjustments were made to the protocol in consultation with representatives from the Blue Ribbon Panel.

Organization of the Protocol

The protocol operates at three levels:

• Articles or reports
• Studies
• Programs

A search of the literature to identify empirically based articles or reports on the programs of interest begins the process. Articles identified for each program are then consolidated into distinct studies, as appropriate. The protocol guides the review of the studies to analyze the quality and findings of the research. Information from the reviews is then summarized to rate the evidence of the programs' effects on student outcomes. See Appendix 1 for the protocol.

Identification of Relevant Articles

AIR conducted a comprehensive literature search to identify published and unpublished materials relevant to the set of programs identified by BEST.[1] An initial pool of 200 candidate programs was determined by nominations of Blue Ribbon panelists, supplemented by BEST staff research, and later narrowed in consultation with AIR. The AIR literature search yielded books and articles from more than 750 education-related journals, unpublished research/technical reports, conference papers from the annual meetings of the American Educational Research Association and other large and small professional educational research organizations, as well as many other "fugitive" materials. In addition, AIR searched the websites of several research organizations (e.g. Rand, The Center for Research on the Education of Students Placed At Risk [CRESPAR] and National Center for Research on Evaluation, Standards, and Student Testing [CRESST]), reviewed program websites and communicated directly with program developers.

The articles identified during the literature search were screened for inclusion in this analysis, that is, which were empirical and which non-empirical? The non-empirical articles were further screened to identify those that compare measurable outcomes for students participating in or completing the program to other students.

Evaluation of Studies

The protocol evaluates the methods, contexts and findings of each study. It also considers:

• the extent of control for selection into programs, and
• the extent of control for other factors that may be offered as alternative explanations for results observed.

To determine the degree of confidence with which claims of program effectiveness can be viewed, the protocol provides guidance for rating the **strength** — or rigor — of research designs. The key to whether studies are of acceptable rigor is the extent to which they reduce the likelihood that the outcomes of the programs for students participating in them can be explained by factors other than the intervention of the program itself, for example, the way in which participants are selected into the program, i.e., selecting gifted or highly motivated students, then attributing their success to the intervention.[2]

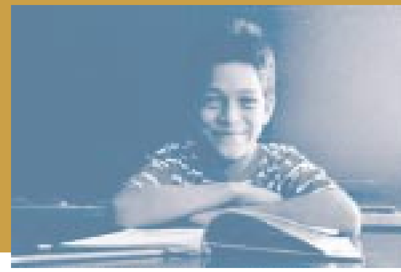
Experiments with randomized assignments and multivariate statistical designs that control for pre-intervention differences represent the clearest examples of acceptable rigor. However, some other research designs may also provide acceptable levels of evidence. This group includes, for example, quasi-experimental designs, in which treatment and non-treatment groups are given pre- and post-treatment tests, then compared. Thus, designs that lack a control group or pretest observations related to the outcome would not pass the protocol. Similarly, multivariate statistical designs that control only for demographic characteristics (e.g., race-ethnicity, socioeconomic status, age, etc.) would not pass the protocol.[3]

In reviewing the context of the study, the protocol considers the independence of the developer from the study and the types of outcomes being assessed (e.g., immediate vs. longer term performance or achievement, attitudinal or other behavioral outcomes). It also considers the generalizability of the study.

The most important step, of course, is evaluating the findings that are related to student outcomes. The protocol takes account of the principal findings reported by the author, the consistency of the reported findings with the data presented in the report, and evidence that the findings are statistically significant differences in the outcome between participants and non-participants. A study is described as having substantially positive findings if an overwhelming number of the results point to positive outcomes for program participants relative to the comparison group, substantially negative if the reverse is true, neutral or mixed if a study meets neither of the preceding criteria.

Assessment of Program Effectiveness

The protocol applies strict criteria in its assessment of a program's effectiveness, depending on the level of rigor with which



it has been studied. At the highest level are those considered **verified**, which have had five studies of acceptable rigor conducted by independent evaluators with substantially positive results. At least two of the studies must be of **high** rigor, and none of studies of acceptable rigor can have substantially negative findings. Taken together, these studies need to demonstrate longer-term as well as immediate effects for one or more of the underrepresented groups.

A program is classified as ***probable*** based on the results of two or more acceptably rigorous studies by independent researchers or evaluators that show positive immediate or longer-term effects and no substantially negative studies. A program is considered ***notable*** if at least one acceptably rigorous study is substantially positive and no studies have substantially negative results.

The second broad category of program effectiveness is based on descriptive evidence of student outcomes. Annual data on student outcomes from developers or from portions of less rigorous evaluations by the developer, a funding agency or independent reviewer may contain descriptive evidence that a program is successfully serving one or more of the targeted populations. Such evidence suggests that the program ***merits further research investment*** on a priority basis.

The third broad category of program effectiveness contains those programs for which sufficient research evidence of program effectiveness is not available, nor was descriptive evidence of student outcomes sufficient to indicate that the program is successfully serving one of the targeted populations. Programs which fell into this category are not described here.

Conclusion

Given the need for on high-quality research to guide policy and funding decisions, we suggest that the BEST/AIR protocol will be useful to researchers, developers and funding agencies. Researchers will have a template against which they can design high-quality studies. Developers will have a template against which they can refine the characteristics of their programs. Funding agencies will have a reason to invest in high quality and effective programs.

REFERENCES

Building Engineering and Science Talent. (2002). *Interim progress report to Congress.* Proceedings from September 26, 2002, hearing, Washington, DC. [Online]. Available: www.bestworkforce.org.

Clewell, B. C., Anderson, B. T., & Thorpe, M. E. (1992). *Breaking the barriers: Helping female and Minority students succeed in mathematics and science*. San Francisco: Jossey-Bass.

Gandara, P., & Bial., D. (2001). *Paving the way to postsecondary education: K-12 intervention programs for underrepresented youth*. Report of the National Postsecondary Education Cooperative Working Group on Access to Postsecondary Education. Washington, DC: National Center for Education Statistics. [On-line]. Available: nces.ed.gov/pubs2001/ 2001205.pdf

Hayward, G. C., Brendes, B. G., Kirst, M. W. & Mazzo, C. (1997). *Higher education outreach programs: A synthesis of evaluations.* [On-line]. Available: www.ucop.edu/sas/publish/pace/index.html

James, D. W. & Jurich. S. (Eds.) (2001). *More things that do make a difference for youth*. Washington, DC: American Youth Policy Forum.

Nicholson, J. J., Weiss, Fl. L. & Campbell, P. (1994). Evaluation in informal science education: Community-based programs. In V. Crane, M. Chen, S. Bitgood, B. Serrell, D. Thompson, H. Nicholson, F. Weiss & P. Campbell (Eds.), *Informal science learning: What research says about television, science museums and community-based projects*. Dedham, MA: Research Communications, Ltd.

Slavin, R. E. & Fashola, O. S. (1998). *Show me the evidence! Proven and promising programs for America's schools.* Thousand Oaks, CA: Corwin Press.

ENDNOTES

1. The search was conducted through a number of electronic bibliographic databases (ERIC, Education Abstracts, Social SciSearch, PsychLit, and dissertation abstracts) dating back to 1980 or earlier.

2. It cannot, however, be presumed that gifted or motivated students will succeed automatically regardless of the quality of the intervention.

3. While studies that control only for demographic characteristics would not be treated as providing research-based support for a program, the results of these studies may nevertheless provide descriptive evidence that program is having a positive impact on underrepresented youth and therefore should be the subject of a more controlled study of program effectiveness.

Chapter 3

Programs that Work

BEST's findings challenge education researchers to implement rigorous research on programs designed to impact students in the math and science pipeline

With the protocol piloted and revised, BEST evaluated the effectiveness of programs that sought to impact the performance of traditionally underrepresented students in the pre-K-12 math and science pipeline. BEST's Blue Ribbon Panel of experts was asked to nominate programs for consideration. From more than 200 nominees, 34 programs were eventually selected for analysis.

Different means were used to cull the list: some programs were too new to have a track record of data or evaluative studies; others focused solely on curriculum development; still others were local versions of the same national program. Time and resources were the limiting factor on the number of programs that could be reviewed in this phase. While representing a diverse collection of programs, this initial review does not presume to be exhaustive of potentially effective programs. By the same token, the sample was large enough to demonstrate the value of the protocol as a rating device whose use could be expanded.

Diversity of Programs Examined

Given BEST's mission, the goal was to review programs that seek to impact each of the various underrepresented groups in the math and science pipeline. The 34 programs represent a variety of educational interventions across the spectrum from pre-kindergarten to 12th grade. Mathematics, science, engineering and technology are the primary subjects on which these programs focused, although several programs also stress reading and two of the programs have a more general college preparatory emphasis.

Some programs specifically target females, students with disability, African Americans, Hispanics or Native Americans. Others are less specific (e.g., targeting all students at risk, urban students, or even all students regardless of race-ethnicity, gender or disability status).

The programs also vary in how the intervention is delivered and the type of developer or sponsor of the program. For example, in addition to core and supplemental classroom materials and activities, many of the interventions involve after-school or outside-of-school activities (e.g., clubs, internships, summer programs or camps, etc.), and two of the programs represent whole school reforms.

Finally, the developers or sponsors of the program include universities and their faculties, government agencies, private organizations and consortia of business, community and educational organizations.

Overview of Results

In reviewing the available research on 34 programs, AIR identified 20 that had research-based or descriptive evidence. In other words, 57 percent of those reviewed have moved beyond the anecdotal stage. Sufficient evidence was unavailable for the remaining programs. Of the nine with research-based evidence:

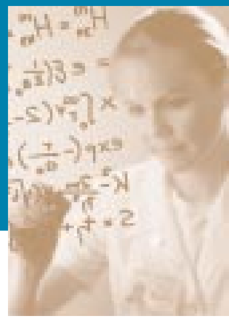
- None had sufficiently rigorous scientific evidence to be considered *verifiably* effective (five or more independent studies with positive findings);
- Two are *probably* effective, according to the criteria of the protocol (at least two rigorous independent studies with positive findings);
- Seven programs are considered *notable* (at least one independent study with positive findings);
- Eleven programs yielded data suggesting that each is *meeting its main objectives* and would benefit from further research.

At first glance, these findings seem disappointing. No programs could be verified as effective, and evidence is thin for 40 percent of the programs reviewed. In some cases, the number of students served by the research-based programs underscores how much remains to be done to solve the challenge of underrepresentation in math and science. These findings do challenge education researchers to implement rigorous research on other programs designed to impact students in the math and science pipeline.

However, these findings also confirm that a diverse set of programs meets the high standards for research-based effectiveness set by BEST's Blue Ribbon Panel. They also indicate that it is possible to move beyond anecdote and intuition to evaluate the effectiveness of programs without setting the bar unrealistically high. And we now have evidence to support efforts to scale up programs that yield evidence of success.

In the remaining sections of this chapter, then, we identify each of the nine programs judged probably effective or notable. We describe its objective(s), highlight its key features and summarize the evidence for its effectiveness. Readers will learn about the set of programs with the strongest research base among those analyzed. One caution: the research base does not indicate which feature or combination of features in any one program account for the positive results.

In Chapter 4, we will extrapolate a set of factors we call "design principles" from the nine programs. These principles are a set of characteristics common to them that can guide and inspire system-wide and targeted programs to help students succeed in their pre-K-12 math and science education.



Programs of Verified Effectiveness

None of the programs evaluated met the criteria.

Programs of Probable Effectiveness

The effectiveness of a program is considered probable when substantially positive results have been found in two or more studies of sufficient scientific rigor. The AIR analysis yielded two such programs: Direct Instruction in Mathematics and Project SEED.

Direct Instruction (in Mathematics)

Direct Instruction (DI) is an approach developed in the late 1960s at the University of Illinois by Siegfried Engelmann and his colleagues. It features carefully specified and sequenced knowledge and skill development for elementary school students in several basic subjects, including arithmetic and mathematics. The DI model aims to provide intense, efficient lessons that will allow all children — even the lowest performing — to master academic skills. The instructional approach applies a "single interpretation" rule to the way in which a teacher presents material to students, based on the assumption that the student will learn the concept that is being presented if there is one, and only one, interpretation available.

The program assigns students to homogeneous groups based on initial testing of their ability and conducts frequent assessments of student progress to maintain grouping by ability. Program materials are commercially available and several organizations contract with schools and districts to assist in implementing the program as part of comprehensive school reform or for use in individual classrooms. Particular emphasis is given to training teachers to deliver instruction in the scripted manner specified by the developer. DI's mathematics programs are used in thousands of schools nationwide, many of which are low-performing schools in high-poverty areas, including both urban and rural districts with enrollments that are predominantly African-American, Hispanic and low-income white students.

Substantial research evidence from rigorous studies comparing DI and non-DI students indicate that DI has an immediate positive effect on mathematics performance of students on standardized tests, although the strongest evidence is based on a large study of students from the early 1970s. Some research also indicates that DI students develop a more positive attitude toward mathematics and their own mathematical abilities, although whether that persists in the long term has not been examined. More limited and less rigorous research indicate that mathematics performance gains of DI students compared to non-DI students are retained into middle school, and that DI students have higher graduation and college admission rates.

Additional rigorous studies conducted by independent researchers are needed to demonstrate that the academic advantages of DI persist throughout high school, perhaps as evidenced by DI students' performance on high school math exams and by their enrollment in advanced math and college preparatory courses. Independent research is also needed to verify the effectiveness of DI in meeting the longer-term goals of promoting underrepresented groups in mathematics and science.

Project SEED (Special Elementary Education for the Disadvantaged)

Project SEED is a supplementary mathematics program begun in the 1960s in Berkeley, California. It targets urban students in grades 3-6, with the aim of improving their critical thinking and problem-solving skills, mathematical achievement levels and academic self-confidence. The project trains professional mathematicians, scientists and engineers to use a Socratic questioning method to teach students abstract mathematical concepts in the classroom. The objective is to have all students be active participants in learning by discovery. The regular classroom teacher observes the instructional practices of the *Project SEED* specialist and attends workshops conducted by the specialists.

Project SEED operates in several large urban school districts serving minority and disadvantaged students.[1] Evaluations of the program in some of these districts found that students who participated in *Project SEED* for a semester outperformed their non-*SEED* counterparts on standardized mathematics tests. Students who received several semesters of *SEED* continued to outperform non-*SEED* students for at least four years and went on to take more higher-level mathematics classes in middle school and high school. While these results reflect the experience of a number of sites in which *SEED* has been adopted, some researchers reporting on studies conducted elsewhere did not find significant differences in performance between *SEED* and non-*SEED* students. Consideration of variation in program implementation across study sites may therefore require more scrutiny as researchers seek to replicate positive outcomes.

Programs of Notable Effectiveness

Seven programs were rated notable because the available research on them includes at least one study of sufficient scientific rigor with substantially positive findings regarding the effectiveness of the program. Most also have descriptive information or less-than-rigorous research evidence. That is, additional research would be necessary to attribute or link specific program characteristics and features to positive results.

Advancement Via Individual Determination (AVID)

AVID is an in-school elective class program established in 1980. Its goal has been to help at-risk middle and high school students who are promising — but underachieving — to undertake and succeed in a college preparatory program, and ultimately enroll in college. The program selects motivated students who enroll in college preparatory classes that include mathematics and science courses. *AVID* provides students with support for their college preparatory classes through tutors and training that focuses on study skills, writing, inquiry, collaborative work groups, and preparation for college entrance and placement exams. Students attend sessions with guest speakers from educational institutions and the business community, and participate in field trips to college campuses or other places of educational interest. Teachers involved in *AVID* attend summer staff development programs and follow-up workshops focused on improving student performance and motivating students from underrepresented groups. The developer reports that the program operates in over 1,500 schools in 21 states and has reached 30,000 students.

Local studies in the San Diego area indicate that a substantial proportion of *AVID* participants are African American or Hispanic. One study showed that *AVID* graduates, particularly Hispanic and African American students, matriculated to college at higher rates than *AVID*-eligible students who did not remain in the program throughout high school. While the research indicates that *AVID* students did not outperform demographically comparable non-*AVID* students in terms of grades and test scores, they did take more college preparatory courses (including algebra in middle school) needed to meet college eligibility requirements than did their non-*AVID* counterparts. Empirical evidence for the program's effectiveness would require specific studies comparing *AVID* students to equally motivated non-*AVID* students who do not receive the same support provided by the *AVID* program.

The Algebra Project

The Algebra Project was developed in the 1960s to provide all students with the conceptual understanding of mathematics necessary to complete algebra successfully and to enter the high school mathematics sequence required for college entrance, in particular to support the transition from arithmetic to algebraic thinking. The curriculum uses activities based on everyday experiences that students understand intuitively to teach basic algebraic concepts. Staff development includes instruction to and support of teachers in implementing the program and a trainer-of-trainers model to build the capacity of local practitioners to take over the professional development aspect of the program. The *Algebra Project* targets inner-city and rural students in the upper elementary/middle school grades, particularly in high-poverty, high-minority communities. A community organization component seeks to build a constituency of parents and other community members to advocate for strong mathematical literacy programs and a peer group support component to increase the self-esteem and self-confidence of students, particularly students of color.

One study revealed a greater improvement in operations with negative numbers for a class of ninth grade, predominantly black and Hispanic, inner-city students introduced to algebra through the *Algebra Project* curriculum in contrast to a younger, gifted and talented class of suburban, predominantly white students with a traditional curriculum. Other research indicated that larger proportions of high school students who had attended *Algebra Project* middle schools were enrolled in college preparatory math courses above introductory algebra, compared with other students in the district. Building the evidence base for the program's effectiveness would require, for example, more extensive comparison of student performance outcomes and more controlled comparison of the longer-term outcomes.

Foundational Approaches in Science Teaching (FAST)

FAST is an inquiry-based science curriculum series for students in grades 6-10 that originated at the University of Hawaii, first implemented in 1971. *FAST* provides a three-year sequence of courses focused on physical science, ecology and the interrelationships among science, technology and society. Working in small teams, students participate in all stages of scientific research: data collection and analysis, literature search and writing reports. For teachers, the program provides 10 days of required training in science content and teaching methods as well as follow-up support for those who are adopting some or all of the curriculum.

The developer reports that over 6,000 teachers in 36 states and 10 foreign countries utilize the program. While *FAST* does not specifically target girls, underrepresented minorities or disabled students, the developer indicates that FAST has been implemented in schools serving diverse student populations including underrepresented minorities, English-language learners and disadvantaged students. Research results indicates that *FAST* students tend to outperform non-*FAST* students on basic thinking in science, laboratory skills and basic science process skills such as graphing and data interpretation. A less rigorous study of longer-term effects found that students enrolled in *FAST* in middle school tended to be more interested in science, had higher grades in high school biology and were more likely than their non-*FAST* counterparts to favor active inquiry over rote learning of biology content. Empirical evidence for the program's effectiveness would require comparative studies that more explicitly sample members of underrepresented groups and by more rigorous study of the longer-term impacts implied by the current research.

Gateway to Higher Education

Gateway, an outreach program of the State University of New York, prepares students in New York City high schools for studies and careers in science, technology, engineering, medicine, health and teaching. The *Gateway* approach is to create an environment of high expectations and academic support, along with college and career awareness. Participants are selected for the program in ninth grade on the basis of prior academic achievement and interest. They attend limited-size classes consisting only of *Gateway* students, and take a four-year program of college preparatory and advanced placement classes in mathematics and laboratory sciences. Students also participate in summer classes, internships in medical and technical settings, and other activities to familiarize them and their parents with postsecondary education and technical careers. *Gateway* students also receive coaching to prepare for college entrance exams. While there does not appear to be any formal staff training and development, teachers for *Gateway* are selected for their teaching qualifications and experience, belief in the students' ability to succeed and a willingness to put in extra time and effort with them.

Beginning in 1986 with 100 freshmen in four New York City high schools, the developer reported in 2002 that about 2,000 students were being served in two *Gateway* schools and *Gateway* programs in nine public high schools. In 1998, the population

participating in the program was 62 percent African American, 19 percent Hispanic, 13 percent Asian and five percent other. Paired student comparisons showed that *Gateway* students had a higher rate of high school graduation, were more likely to take end-of-course examinations in mathematics and science subjects, as well as SAT tests, and to achieve higher SAT scores than their non-*Gateway* counterparts. Descriptive evidence indicates that more *Gateway* students enroll in college than the state and national averages. Over three-quarters of the students who entered the program during its first five years, and 92 percent of those who graduated from the program had gone on to college. Over half of respondents to a follow-up survey of college-goers from *Gateway's* first three graduating classes reported that they had remained in science-based fields including medicine, health, engineering and computer science. Because admission to *Gateway* is selective, building the evidence for the program's intermediate and long-term effectiveness would require, for example, extending the paired comparisons used to examine immediate effects.

Project GRAD (Graduation Really Achieves Dreams)

Project *GRAD* is a K-12 school reform program that began in 1989 as a scholarship program for students in a low-income, high-minority high school. Rather than specifically promoting students in mathematics and science-related fields, the program aims more fundamentally at reducing drop-out rates and increasing the rate at which the students attend college.

GRAD has developed a comprehensive approach to addressing these objectives by utilizing school feeder patterns to implement a series of program interventions. The interventions are designed to improve classroom management and discipline, proficiency of elementary and middle school students in reading and mathematics, and parent and community involvement in addressing issues that adversely affect the success of children in school. The high school component of the program informs parents of college scholarships and recruits incoming freshman to a scholarship program in which students commit to taking college preparatory classes, maintain a minimum grade point average, attend summer institutes at a local university to become familiar with university expectations, and graduate from high school on time. In exchange they are eligible for a four-year undergraduate scholarship. Professional development of teachers includes training in the reading and mathematics intervention programs and in a program on classroom management and discipline.

The project serves primarily low-income, inner-city schools with high proportions of African American and/or Hispanic students. By 2002 Project *GRAD* had been adopted in five feeder school systems in Houston, where it was first initiated, and in eight other large urban areas nationally. In a moderately rigorous study from Houston, several years of statistically controlled comparisons indicate that Project *GRAD* students outperformed students in non-*GRAD* schools on standardized reading and mathematics tests. The study also reported an increase in the rate of college-going among graduating high school students.

Comparative investigation of the program's effectiveness needs to be expanded to include other sites, address the original goals of the program and assess the extent to which the program results in more underrepresented minorities pursuing mathematics and science related fields of study. An independent evaluator is currently conducting a study at several other elementary and secondary sites. However, longitudinal studies comparing reductions in dropout rates and increases in college enrollment rates of Project *GRAD* and non-*GRAD* students are needed as these represent the ultimate goals of the program. From BEST's perspective, such studies also need to assess the extent to which Project *GRAD* students develop an interest in mathematics, science and engineering, and go on to pursue college majors and careers in those fields.

Puente

Puente is a California-based college access program designed to provide a bridge for Latino students from high school to college and from community colleges to four-year colleges. The program at both the high school and community college levels emphasizes instruction in critical thinking and writing, intensive counseling to prepare students for college or transfer from community college and mentoring of students to introduce them to opportunities they might not otherwise envision. The program selects English-speaking Latino students who are motivated to improve their academic performance and participate in a college preparatory curriculum. Participants enroll in a two-year college preparatory English class in ninth and 10th grade. The program does not focus explicitly on BEST's targeted fields of science, technology, engineering and mathematics. However, it explicitly aims to increase Latino student enrollment in college. The program also counsels students and parents on college preparatory requirements and progress, and engages the Latino community in the college entry process with Latino professionals. *Puente* operates in about 36 high schools and 54 community colleges in California.

The results of a longitudinal study of matched samples of students found that *Puente* students were more likely than non-*Puente* students to matriculate to a four-year college, which is consistent with the major objective of this program. While *Puente* students were more likely to have completed college entrance examinations, completion of college preparatory courses was more prevalent for *Puente* as compared to non-*Puente* students only among the higher achieving students. A follow-up analysis two years after high school graduation indicated that *Puente* students were more likely to be enrolled in some form of college than non-*Puente* students. Comparing the course-taking patterns and college majors of matched samples of *Puente* and non-*Puente* students to see what impact the program has on recruiting Latino students into mathematics and science-related fields would contribute to the evidence for the program's effectiveness.

Yup'ik Mathematics

This program seeks to improve the mathematical thinking and performance of Yup'ik students at the elementary school level while reinforcing Yup'ik culture. The program stems from the voluntary collaboration, beginning in 1998, among tribal elders, bilingual aides and university faculty to develop curriculum in mathematics. A hands-on approach to learning that links intuitive understanding to formal mathematical reasoning through use of everyday, practical knowledge is used.

Yup'ik Eskimos are primarily located in rural villages and urban areas of southwestern Alaska. While the program has only begun to be implemented, a controlled study showed that students in randomly selected classrooms in rural and urban school districts who were exposed to a three-week culturally based module for teaching several geometric principles demonstrated greater improvement in understanding those principles compared with a control group of students who were taught using a standard textbook. As formalization of the curriculum is recent and ongoing, more work is needed before even the immediate effects of the program can be completely evaluated, let alone longer-term effects. While *Yup'ik Mathematics* may serve a small population, the initial finding that students exposed to culturally relevant curriculum show greater improvement in mathematics performance make this a noteworthy program.

Programs that Merit Further Research Investment

In addition to the nine programs already described, the search for exemplary programs also yielded 11 additional efforts whose descriptive evidence is sufficiently robust to merit further research on a priority basis. These programs are:

- American Chemical Society (ACS) Project SEED
- Detroit Area Pre-College Engineering Program (DAPCEP)
- Disabilities, Opportunities, Internetworking and Technology (Do-It)
- El Paso Collaborative for Academic Excellence (EPCAE)
- Family Math
- MATHCOUNTS
- Mathematics, Engineering, Science Achievement (MESA)
- Operation SMART – Girls Inc.
- Texas Pre-freshman Engineering Program (TexPrep)
- University of North Carolina Mathematics and Science Education Network (MSEN)
- Xavier University Summer Science Academy

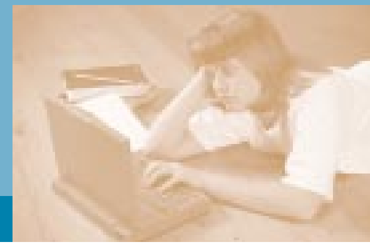
These programs are profiled in the User's Guide at the end of this report, along with fuller descriptions of those interventions that met the criteria of the research protocol. Sufficient descriptive evidence was not available on 14 of the 34 programs evaluated in the BEST/AIR review.

Finally, it is important to recognize that some programs that are ramping up may in the future produce evidence that meets the criteria of the protocol. For example, Materials World Modules (MWM), an NSF-funded supplementary science-engineering program for students in grades seven to 12, conducted a feasibility study in 2001-2002 and is currently conducting a nationwide study of student achievement. The external evaluator reports that preliminary results for a study to be reported on in the fall of 2004 look positive. The U.S. Department of Defense is using the program in 10 overseas high schools and is planning to extend implementation to schools stateside.

Even this brief introduction to the programs with demonstrated effectiveness suggests that they may share characteristics to which their success may be attributed. The BEST/AIR review found that, indeed, there are some shared principles on which these programs are built, and from which other efforts can benefit. The next chapter discusses these principles in greater detail. See Appendix 2 for greater program details.

ENDNOTE

1. For example, Project Seed operates in Dallas; Detroit; Philadelphia; Camden, New Jersey; Indianapolis; Milwaukee and the San Francisco Bay Area.



Chapter 4

Implications for Policy and Practice

BEST's design principles suggest a direction for educational policy makers who understand the critical importance of deepening and broadening the American technological talent base

This report has outlined the development and application of a rigorous, research-based protocol to evaluate programs that work for students underrepresented in technical fields. The protocol may also be used as an instrument to maintain an ongoing process of examination of targeted and system-wide educational initiatives the protocol provides a solid foundation upon which to create an authoritative users' guide to effective programs.

This chapter presents a set of design principles extracted from programs that met the criteria of the protocol and a complementary set of policy imperatives to broaden the base of pre-K-12 math and science education.

Design Principles

BEST identified a framework of design principles or shared features extracted from programs with research-based evidence of effectiveness. These design principles are not intended as a causal explanation of what works, yet are derived as reasonable inferences. They are drawn from a small and varied sample of programs, none of which met the standard of evidence that could be deemed "verified." Nevertheless, they are a first approximation of what it takes to succeed over time in a demanding environment.

Applying the definition that design principles are usable by all people, to the greatest extent possible, without the need for adaptation or specialized design, BEST extracted the following five common features:

- ***Defined outcomes*** drive the intervention and are successfully accomplished for the entire target population. Students and educational staff agree on goals and desired outcomes. Success is measured against intended results. Outcome data provide both quantitative and qualitative information. Disaggregated outcomes provide a basis for research and continuous improvement.
- ***Sustained commitment*** enables effective interventions to take hold, produce results and adapt to changing circumstances. Its components are proactive leadership, sufficient resources and steadfastness in the face of setbacks. The minimum conditions for assuring sustained commitment are continuity of funding and of support at the individual school and school district levels.
- ***Personalization*** acknowledges that the goal of intervention is the development of students as individuals. Student-centered teaching and learning methods are core approaches. Mentoring, tutoring and peer interaction are integral parts of the learning environment. Individual differences, uniqueness and diversity are recognized and honored.
- ***Challenging content*** provides the foundation of knowledge and skills that students master. Curriculum is clearly defined and understood. Content is related to real-world applications, goes beyond minimum competencies, and reflects local, state and national standards. Students understand the link between the rigor of the content they study and the career opportunities which await them later in life. Appropriate academic remediation is readily available.
- ***Engaged adults*** who believe in the potential of all students provide support, stimulate interest and create expectations that are fundamental to the intervention. Educators play multiple roles as teachers, coaches, mentors, tutors and counselors. Teachers develop and maintain quality interactions with students and each other. Active family support is sought and established.

While these design principles do not account for the success of any particular intervention, they represent a checklist of the essential characteristics of programs that have passed through a research filter. Viewed as a whole, they yield a number of insights that have the potential to inform policy as well as practice:

- The design principles of programs that work appear to comprise a package rather than an á la carte menu.
- The fundamentals associated with programs that target students from underrepresented groups seem apt to work for all students.
- The design principles derived from math and science programs appear to be applicable to other disciplines.

In a nutshell, these design principles provide a framework for bridging the microanalysis of effective programs and the macro-level requirement of widespread increased math and science readiness. Such bridging will be essential to capitalize on the strengths of targeted programs in order to make the system as a whole more responsive to all students. Finally, these design principles give the developers of existing and new programs a framework for thinking about their own programs. Yet they go beyond being a mere checklist for program developers and evaluators. On a structural level, these design principles also suggest a direction for educational policy makers who understand the critical importance of deepening and broadening our technological talent base.

Policy Imperatives

The program evaluation conducted by BEST and AIR yielded three policy imperatives: first, to deepen the knowledge base regarding what works to prepare and interest students who are underrepresented in math and science; second, to tighten the link between knowledge, policy and practice in this area; third, to align targeted programs and system-wide approaches more closely. The remainder of this chapter outlines these imperatives and proposes a course of action to address each.

1. Deepen the Knowledge Base

The fact that a substantial search for effective programs turned up relatively little research evidence is itself an important finding. The dearth of evidence, corroborating the work of others, is by no means limited to math and science education. Its causes are widely recognized and mutually reinforcing. As noted in Chapter 2, rigorous third-party evaluation is expensive, especially for programs operating on "soft monies" at the periphery of the pre-K-12 system. At the same time, such evaluation is rarely a condition of starting up or sustaining an intervention. Research evidence is useful, but has not been a "must." As a result, program developers have generally looked for the most accessible and lowest cost measures of effectiveness they can find. This approach has yielded little upon which to build.

The national interest lies in deepening the research base as quickly and cost effectively as possible. While a firmer understanding of what works is no magic bullet, it could improve the allocation of resources, support more informed decisions by students, parents and educators, and provide the foundation for vitally needed continuous improvement to meet a difficult and complex challenge.

The time horizon for rigorous evaluation is both short-term and long-term. The need to better prepare the underrepresented majority in math and science is ***immediate***. A realistic, well-focused research agenda on this issue in pre-K-12 math and science is needed. Based on lessons learned from this report such an agenda should include:

- Using the BEST/AIR evaluation protocol to inventory the research on math and science interventions that focus on the groups underrepresented in these disciplines.
- Supporting this research inventory with partnerships among congressional, federal, state, local and private sector interests.

The U.S. Congress, federal agencies, science- and engineering-based industries and philanthropies that fund math and science programs targeting underrepresented groups share the challenge to provide leadership in this effort.

2. Tighten the Links among Research, Policy and Practice

Deepening the knowledge base clearly is not an end in itself. There is much more to be gained from a sustained program of improved research into how and why certain innovations in math and science are more effective than others. Most importantly, the knowledge gained can give policy makers and practitioners faster, better, more relevant information on which to base their decisions, and ensure that it is grounded in sound empirical research. Since operational needs are short-term, it is essential to be able to extract value from the existing body of research or risk irrelevance.

Much needs to be done to integrate policy, practice and research more fully. The creation of a high-profile research capability within the U.S. Department of Education is a recent innovation whose long-term impact has yet to be felt. For now, exchanges among the members of the pre-K-12 BEST Blue Ribbon Panel suggest that practitioners rely mainly on their own professional judgment or that of respected peers in deciding whether to adopt — or how to adapt — new math and science programs. Until research findings are easily accessible and seen to matter, they are not likely to draw the attention of a harried superintendent or principal.

One way to tighten links is to make sure that research findings are readily available to and easily understood by practitioners and policy makers. The purpose of the User's Guide in the Appendix 2 to this report is to begin developing a research-based communications tool that draws bottom-line conclusions regarding what works.

The same considerations hold for many policy makers, particularly members of Congress. BEST recommends that congressional hearings on the interrelationship between policy, practice and research in pre-K-12 math and science education focused on pipeline issues could create a national agenda for connecting the three domains more closely, and secure their funding.

3. Align Targeted Programs and System-wide Approaches

The programs evaluated in this report were all established to meet focused needs, whereas pre-K-12 math and science education has the mission of meeting broader needs. What is most striking, however, is that the design principles extracted from programs that work are relevant to reforms that address entire education systems:

- Whereas defined outcomes at the programmatic level refer to measurable results achieved by individuals (usually a limited number), they encompass the performance of whole schools at the system-wide level. But the basic value of establishing performance measures in relation to well-conceived and clearly understood goals is the same.
- While sustained commitment for programs means the continuity of enabling factors, such as leadership and resources, that lead to success, the same continuity is equally indispensable and just as difficult to achieve system wide. In both cases, it is the quality, consistency and steadfastness of leadership that spells the difference.
- Personalization is a not only a bedrock feature of effective programs, but also the design principle that underpins system-wide efforts to reduce class and school size, for example. The objective at both levels is to reach individual learners.
- Although programs that work are defined by their

challenging content, often tailored to the needs of specific populations, such content must align with system-wide standards to insure that students are prepared to take the next educational step in curricular content.

- Engaged adults are a signature feature of effective programs, and they are equally vital to system-wide success. The difference is that programs often attract qualified and committed teachers, while a host of forces at the system-wide level skew teaching resources away from underrepresented groups. For example, it is at the system level where decisions are made to rationalize out-of-field teaching assignments, a problem endemic for too many teachers of science and math at the elementary and middle school levels, and at science and math entry-level courses in high school.

These illustrations point up the need for a more comprehensive dialogue between programs that are succeeding "in the trenches" and decision-makers with system-wide objectives and responsibilities. A serious dialogue can contribute to a national solution in two important ways: first, by developing an action plan to insure that the programs that work do not remain at the margins of the system; and second, by increasing the capacity of pre-K-12 math and science education as a whole to develop the talent of students from all groups.

Effective programs have what it takes and the system needs more of what it takes. The talent imperative demands that we build on the strengths of both.

Appendix 1

BEST Pre-K-12 Research Literature Review Protocol For Pilot Study

The protocol operates at three levels:

- Programs
- Studies
- Articles or reports

The protocol begins with a search of the literature to identify empirically based articles or reports on the programs of interest. The articles identified for each program are then consolidated into distinct studies. The protocol guides the review of the studies that have been identified for each program (e.g., quality and findings of each individual study). The information from the reviews is used to summarize and rate the evidence of the programs' effects on student outcomes.

The steps in the protocol that are undertaken for each program include:

Step 1: Identification of program articles and reports
Step 2: Screen for empirically based articles and reports
Step 3: Screen for comparative articles and reports of student outcomes
Step 4: Consolidation of articles into distinct studies
Step 5: Review of study methodology
Step 6: Summary of study context
Step 7: Summary of study findings
Step 8: Summary rating of evidence of effects of program
Step 9: Description of the evidence of effects

PROGRAM/INTERVENTION COVER PAGE

Program/Intervention Name: ______________________________

Alias: ______________________________
(e.g., names that may be associated with earlier versions or components of the program/intervention)

Developer(s): ______________________________
Name: ______________________________
Address: ______________________________
Telephone: ______________________________
Email: ______________________________

Year first implemented in an educational setting: ______________
Where implemented: ______________________________
Major reformulations of program/intervention (if available)
Year: ______________________________
Description: ______________________________
Where reforms were implemented: ______________________
____

STEP 1: IDENTIFICATION OF PROGRAM ARTICLES AND REPORTS

Objective: Comprehensive identification of articles/reports related to program.

The literature search is conducted through electronic searches of a number of bibliographic databases, including:

- ERIC
- Education Abstracts
- Social SciSearch
- PsychLit
- dissertation abstracts dating back to 1980 or earlier.

In addition to referencing published books and articles from 750 or more education-related journals, the bibliographic databases used in this literature search include unpublished research/technical reports, conference papers from the annual meetings of the American Educational Research Association and other large and small professional educational research organizations, and many other "fugitive" materials.

The contents of program websites where available are also examined, as are the websites for several research organizations (e.g. Rand, CRESPAR, and CRESST). Assistance in identifying published and unpublished reports is solicited from each of the developers. Developers are sent an e-mail requesting unpublished as well as published evaluations of their programs. A follow-up request is made by telephone to developers who do not reply to the e-mail request.

STEP 2: SELECTION OF EMPIRICALLY BASED ARTICLES AND REPORTS

Objective: Distinguish empirical reports and from purely descriptive reports.

Reports are screened to identify those that contain statistics pertaining to the program. This is a gross screening to separate out promotional articles and purely descriptive reports that provide no potentially analyzable outcome data. This does not screen out "secondary" or review studies where the results of original studies are simply being reported or are being assessed by a third party. The latter are carried through to the next stage so that we can determine whether we have identified the primary or original study.

STEP 3: SCREEN FOR COMPARATIVE ARTICLES OF STUDENT OUTCOMES

Objective: Select articles based on studies that meet four criteria:

- designed to address specific research questions regarding the effectiveness of the program;
- contain a comparison for students participating in the program;
- examine student outcomes; and
- include information about the context in which a program is carried out.

Study design

Yes No
❑ ❑ Does article/report address a specific research question or set of questions regarding the effectiveness of the program (e.g., annual reports that provide statistical information about a program do not qualify as studies, unless they contain a research design aimed at addressing identifiable research questions pertaining to the effectiveness of the program)?

Yes No
❑ ❑ Outcomes for subjects participating in or completing the program are compared to outcomes for subjects not participating in, or not completing the program.

Yes No
❑ ❑ Outcomes for subjects participating in the program are compared to outcomes for an extant reference group (includes state or district averages, etc.).

Comments:____
____
____

(Note: This is a general screen. The nature of the comparison or control group will be further evaluated in Step 5, review of methodology.)

Contextual information

Yes No
❑ ❑ At least minimal contextual information describing activities related to the program (e.g., teachers participated in related professional development activities, students participated in weekly mentoring sessions)

Comments:____
____
____

Effects

Yes No
❑ ❑ Study contains quantitative analysis of measurable student outcomes?

Type of outcomes reported (check all that apply):
❑ Student achievement (e.g., test scores, promotion rates, grades, college enrollment)
❑ Other student outcomes (e.g., course enrollment, college major, post-college job choice, science fair participation, science club participation, career interest and attitudes)
❑ Although not used to screen, check if the study analyzes other outcomes?
❑ Structural or environmental outcomes (e.g., changes in teachers' instruction, school climate)
❑ Process outcomes (e.g., number of students participating in the program)

Comments:____
____
____

Decision rules

To pass this initial screen, the study reported in the article should meet the following criteria:

- Addresses identifiable research question pertaining to effectiveness of the program.
- Uses a comparison group, either selected a priori or an extant reference group.
- Describes the context.
- Includes quantitative analysis of student outcomes.

Step 3 summary

Initial Screening:
- Accepted
- Rejected: If so, why: ________________________________
____

Note: If an article is rejected because it does not contain a comparison group, but it does include quantitative longitudinal data of student outcomes for one or more cohorts, assessments will be made regarding:

- acceptable methodological rigor of study (i.e., adequate sampling design, high response and retention rate, and reliable and valid student outcome measures);
- availability of comparable data from extant data sources (i.e., extant federal data sources with comparable outcome data that can be used to identify a reasonably statistically controlled comparison based on social and demographic characteristics).

If the study on which the article is based is of acceptable methodological rigor and a reasonable comparison group can be identified in the extant data, then the context of the study will be summarized (i.e., see Step 6) and a secondary analysis will be conducted by comparing findings for the study group to a comparison group obtained from the extant data. This type of study will be classified as acceptable albeit with less rigor (i.e. one based on a non-equivalent or statistically controlled comparison group). The findings of these analyses will be recorded in Step #7 and considered in the rating of evidence for the program in Step #8.

STEP 4: CONSOLIDATION OF ARTICLES INTO DISTINCT STUDIES

Objective: Determine number of unique studies and identify article(s) to be reviewed to capture the empirical findings for each study.

Articles based on the same treatment group are considered to belong to the same study. For each study:

Name of study: ____

Conducted by: ____

Date(s) conducted: _____________________________________

Study acronym (to be assigned by reviewer): _________________

Comments: ____

List empirical articles/reports associated with study:

Article/ Report #	Author(s)	Title and Reference (i.e., include journal, pages, etc.)	Review Y/N
1			
2			
3			
4			
5			

Check the articles for which findings will be summarized (i.e., one article may be a comprehensive report of findings from a number of articles, or may be a final report that supplants the preliminary findings reported in earlier articles.)

STEP 5: REVIEW OF STUDY METHODOLOGY

Objective: Evaluate the rigor of the study based on the quality of its design, data and measures.

Type of study

Yes ❑ No ❑ Experimental (i.e., a priori random or matched assignment to treatment and control groups)

Yes ❑ No ❑ Multivariate analysis with statistical controls

Yes ❑ No ❑ Quasi-experimental (e.g., pre-test and posttest with non-equivalent comparison group, cohort comparison, static group comparison)

Yes ❑ No ❑ Other, specify: ________________________________
____
____

Design controls – Check all that apply

❑ A priori random or matched assignment to treatment and control groups
❑ Pre-test and post-test measures on treatment and control group
❑ Cohort comparison
❑ Static group comparison (i.e., no pre-test/post-test comparison)
❑ Extant reference group comparison
❑ Other, specify

Statistical Controls – Check all that apply

❑ Pre-program differences theoretically linked to outcomes examined in study
❑ Available pre-program measures (e.g., performance, behaviors, and abilities) correlated with outcomes examined in study
❑ Demographic characteristics (e.g., race-ethnicity, gender, age etc. of students)
❑ Other, specify: ______________________________________
____
____

Comments:____
____
____

NOTE: The design and statistical controls are not exhaustive of all possible ways that a study may be designed.

Data

Yes No

❑ ❑ Is the data analysis longitudinal (i.e., outcome data on same cohort of subjects for three or more years after treatment initiated)?

Comments:____

____

____

Measures

Yes No

❑ ❑ Is at least one of the outcome measure(s) valid and reliable?

If yes, what is basis for conclusion (check all that apply):

❑ study uses nationally or well-established instruments (e.g., SAT-9, DRT) developed independently of the program and study.

❑ study uses extant measures (e.g. GPA, course enrollment, college major, post-college job choice).

❑ study effectively assesses the reliability and validity of out come instruments developed specifically for that program or study.

If no, explain basis of conclusions ______________________

____

____

Other comments:_______________________________________

____

____

Decision rules

1) Studies must have at least one outcome measure that can be considered to be reliable and valid.
2) The following categories guide the rating of the research-based evidence of each study:

Classification	Characterization of study
Most rigorous	• Experimental with a priori randomization or matching of subjects • Multivariate longitudinal study with statistical controls for pre-program differences and other variables theoretically related to outcomes examined in study
Modest rigor	• Multivariate study with statistical controls for available preprogram measures correlated with outcomes examined in study • Quasi-experimental with pre-test and post-test on treatment and non-equivalent comparison group or cohort comparisons
Low rigor	• Reference group comparison with demographic controls only • Multivariate study with less stringent controls • Static group comparisons
None	• Study presents only gross comparisons without controlling treatment and comparison/reference group characteristics.

Studies rated modest or higher are considered to be of acceptable rigor. Although studies rated low are not considered to be of acceptable rigor to qualify as research-based evidence, the findings will be recorded as they may provide descriptive evidence that suggest a program is effective and would benefit from further study.

3) Note that studies are not required to have a longitudinal design (e.g., prospective with repeated measures, retrospective) or contain longitudinal data to be considered rigorous or acceptable. However, within each of the above levels, the availability of longitudinal data would be considered a desirable feature.

Summary rating of study methodology

Methodology screening:

• Accepted: If so specify level of rigor (e.g., most rigorous or modest)

____

____

• Rejected: If so, why:

____

____

STEP 6: SUMMARY OF CONTEXT OF STUDY

Objective: Specify the independence of the study vis-à-vis the program developer; describe the focus of the outcomes analyzed in the study; depict the generalizability of the study based on its setting, location and characteristics of the sample being studied; and identify the theoretical and practical contributions of articles based on this study.

Independence of the study

Yes No

❑ ❑ Study conducted by developer?

❑ ❑ Study conducted by evaluator commissioned by developer or funding agency?

❑ ❑ Study conducted by independent researcher?

Comments:____

____

Focus of outcomes

Articles based on study examine outcomes specific to (check all that apply):

Yes No
- ❑ ❑ Math
- ❑ ❑ Science
- ❑ ❑ Technology
- ❑ ❑ Engineering
- ❑ ❑ General outcomes not specific to any of the above disciplines

Comments:____
____
____

Articles based on this study assess:

Yes No
- ❑ ❑ Overall program effects
- ❑ ❑ Design features or program component effects

Comments:____
____
____

The student outcomes examined in articles based on this study include:

Yes No
- ❑ ❑ immediate effects on students (i.e., while in program or at same grade level as when in program)

  If yes, check all of the outcome categories that apply:
  - ❑ performance/achievement (e.g., grades, test scores)
  - ❑ other behavior (e.g., course selection, school activities)
  - ❑ attitude (e.g., interest in attending college, careers in science)
  - ❑ other, specify

- ❑ ❑ longer-term effects (i.e., intermediate and long range effects are measured at the grade level following completion of program or later.)

  If yes, check all of the outcome categories that apply:
  - ❑ performance/achievement (e.g., grades, test scores, college entrance or graduation)
  - ❑ other behavior (e.g., course selection, college major, school activities)
  - ❑ attitude (e.g., interest in attending college, careers in science)
  - ❑ other, specify

Comments:____
____
____

Generalizability of the study

Setting(s) of study:

Yes No
- ❑ ❑ Urban
- ❑ ❑ Suburban
- ❑ ❑ Rural
- ❑ ❑ Information not reported

Geographic location(s) include:

Yes No
- ❑ ❑ South
- ❑ ❑ North
- ❑ ❑ East
- ❑ ❑ West
- ❑ ❑ Midwest
- ❑ ❑ U.S. (i.e., nationally representative study)
- ❑ ❑ Outside of U.S.

Comments:____
____
____

Characteristics of the sample

Sample unit(s) in study include (check all that apply):

Yes No
- ❑ ❑ Students
- ❑ ❑ Teachers
- ❑ ❑ Administrators
- ❑ ❑ Schools
- ❑ ❑ Other, specify: ______________________________________

____

For student samples:
Sample size N= ______________
Percentage female ______________
Percentage
Am. Indian\Alaska Native ______________
Asian\Pacific Islander ______________
Black\African American ______________
Hispanic\Latino ______________
White ______________
Total 100%

Percentage students with disabilities (describe):________________
____
____

Average SES:____

Grade level(s) or age(s):________________________________

Comments:____
____
____

Description of theoretical framework explaining the relationship between the program and the outcomes:
____
____

Provide assessment of implementation process that may be useful to practitioners:
____
____

STEP 7: SUMMARY OF STUDY FINDINGS

Objective: Compile the findings for each study and assess what the study demonstrate about the student outcome effects of the program.

Unless the review of methodology indicates that a study is completely unacceptable, fill in the attached with the following information about the findings of a study located in the articles reviewed for that study.

In order to be identified as a finding for purposes of this review, each purported result from a study must:

- relate to a student outcome;
- be identified by the author of the report (i.e., articulated) as a finding;
- be consistent with the data as presented in the report;
- include test of statistical significance or data needed to calculate tests of significance and effect sizes;
- be distinguishable from other findings (i.e., add value to our understanding of the impact or lack of impact of a program).

The contents of each row are described in Table 1 page 37.

Summary of reported findings for each study

The matrix of potential student outcomes covered in this review includes a time dimension (i.e., immediate vs. longer-term) and a substantive dimension which we have divided into categories of performance or achievement, other behaviors, attitudes and other outcomes. The result is eight potential arenas of student outcomes (see Table 2 on page 38). A summary of the implications of a study for program effects should indicate the arenas of outcomes being assessed as few studies are likely to focus on all arenas.

Studies may have a mix of positive, negative and neutral findings, making it difficult to summarize the findings succinctly. The following decision rules are intended to help the reviewer make a summary statement about the overall direction of findings of the study.

For a study to be identified as demonstrating ***substantially positive*** effects over 67% of the principal findings that are consistent with the data must be statistically significant at the $p<.05$ level or have effect sizes of .25 or larger.

For a study to be identified as having ***substantially negative*** effects, over 67% of the principal findings that are consistent with the data must be significantly negative with effect sizes of .25 or larger.

A study that meets none of these criteria would be identified as ***neutral*** or ***mixed***.

STEP 8: RATING OF EVIDENCE OF EFFECTS OF PROGRAM

Research-based evidence: verified or substantiated

For program effectiveness to be considered verified or substantiated, a program must have a research base that includes all of the following:

- At least five acceptable studies conducted by independent researchers or evaluators that have substantially positive findings, of which at least two experimental studies with random assignment or multivariate studies with strong theoretically based controls for pre-program differences.
- Taken together the studies show positive immediate AND longer-term effects.
- None of the studies with an acceptable methodological rigor (i.e., rigorous or less than rigorous, but acceptable) are rated as having substantially negative findings.
- At least one study that contributes to theory building and/or implementation.

Table 1. Information to Record in Table of Findings*		
Item	**Characteristic**	**Description**
1	Finding #	Sequential ID# assigned to each finding
2	Study ID	Use study ID assigned in Step 4
3	Article ID	Use article/report # assigned in Step 4
4	Page(s)	Page(s) on which finding is reported
5	Unit of Analysis	Student, class, achool, grade
6	Treatment Group	List defining characteristics of treatment group.
7	Comparison Group	Describe the comparison group or extant reference group, including matching criteria (e.g., matched on SES).
8	Dimension	Note the dimension measured (e.g., achievement gains on biology subtest, graduation from high school).
9	Study Design pertaining to this finding	Specify whether study is rigorous-experimental; rigorous multivariate; less than rigorous but acceptable –quasi-experimental; etc.
10	Measurement Instrument	Where applicable, name the measurement instrument (e.g., Stanford 9). Note whether or not the instrument was developed locally or exclusively for this research or program.
11	Timing of Outcome	Immediate or longer term outcome
12	Type of Outcome	Performance/achievement, other behavior, attitude, or other
13	Finding	Briefly describe the direction and nature of the finding (e.g., test scores increased over one year) for which tests of statistical significance have been conducted.
14	Outcome	Finding of positive, negative, or no program effect
15	Accuracy of Interpretation	Finding agreed with data presented, disagreed with data, or it is unclear whether finding is supported by the data.
16	Statistically Significant	Yes/No/Not reported
17	Level of Significance	If reported, at what level
18	Statistic Reported	Means, proportions, correlations, etc.
19	Size of Treatment Group	N for treatment group
20	Treatment Group Outcome	Mean, proportion or correlation for treatment group
21	Std dev. –Treatment	Standard deviation for treatment group
22	Size of Comparison Group	N for comparison group
23	Comparison Group Outcome	Mean, proportion or correlation for comparison group
24	Std dev. - Control	Standard deviation for control group
25	Overall Standard Deviation	Pooled standard deviation for treatment and control group
26	Effect Size	
27	Comment	

*Items entered as available in or calculable (i.e., effect size) from information reported in study.

Table 2. Summary Rating of Findings of Study for Program Effects				
Student Outcome	**Outcome Assessed (Yes/No)**		**If Assessed:**	
		Substantially Positive	**Substantially Negative**	**Neutral or Mixed**
Immediate Effects				
Performance/achievement				
Other behavior				
Attitudes				
Other				
Longer-term Effects				
Performance/achievement				
Other behavior				
Attitudes				
Other				

- The acceptable studies taken together indicate substantial positive findings in a range of situations or substantial positive outcomes for one or more of the underrepresented groups (i.e., girls, underrepresented minorities, students with disabilities).

Reasearch-based evidence: Probable

For program effectiveness to be considered probable, a program must have a research base that includes all of the following:

- At least one rigorous (i.e. experimental or multivariate) study by an independent researcher or evaluator, or two less than rigorous, but acceptable studies by an independent researcher or evaluator with substantially positive findings.
- Taken together the studies show positive immediate or longer -term effects.
- None of the studies that meet minimal standards (i.e., rigorous or less than rigorous, but acceptable) are rated as having substantially negative findings.
- At least one study contributes to theory building and/or implementation.

Reasearch-based evidence: Notable

For a program to be considered notable, a program must have a research base that includes all of the following:

- At least one acceptable study that has substantially positive immediate or longer term effects.
- None of the studies that meet minimal standards are rated as having substantially negative findings.
- At least one study contributes to theory building and/or implementation.

Descriptive evidence: Benefit from further research

To be considered a program that could benefit from further research, the combination of less than rigorous research and descriptive data needs to demonstrate that a program is meeting its intended objective even though rigorous research-based evidence is not available to permit a controlled comparative test of the program's effectiveness.

Research data unavailable

Sufficient research data or descriptive evidence of a program's effectiveness are not available.

STEP 9: DESCRIPTION OF PROGRAM EFFECTS

This text should include the following types of information from the research-based and descriptive evidence.

Research-based evidence:

- The research base on this program includes (n) studies of the most rigor, (n) of modest and (n) of low rigor.
- Collectively, the research studies looked at the program in the following types of geographic settings (e.g., midwest, south, north, east, west; urban, suburban, rural, mixed) and educational settings (e.g., high poverty schools).
- Collectively, the research studies examined effects for approximately (n) students, recognizing that there may be some overlap in samples.
- Collectively, these studies provide information on the effects of the program on the following key subgroups (i.e., students who are female, underrepresented minorities, or with disabilities) in the following key subject areas (i.e., science, mathematics, engineering, technology).
- The greatest strength of the outcomes research base for this program is…
- The greatest weakness of the outcomes research base for this program is…

Descriptive evidence:

Data showing that the program is serving its target population and meeting its intended objective (e.g., high proportion of students taking college preparatory math curriculum or attending college).

Findings:

Indicate whether or not the research evidence permits one to conclude that the program is verifiably, probably, or potentially effective, and, if so, for what subgroups.

Where there is not adequate research-based evidence for the effectiveness of a program, indicate whether the combination of research-based and descriptive evidence is sufficient to recommend a program for rigorous research.