Finding out what works
Agency Efforts to Strengthen the Evaluation of Federal STEM Education Programs

A Report of the National Science and Technology Council
Education Subcommittee

In Response to Recommendations of The
U.S. Department of Education Report of the Academic Competitiveness Council

May 16, 2008

About the National Science and Technology Council

The National Science and Technology Council (NSTC) was established by Executive Order on November 23, 1993. This Cabinet-level Council is the principal means within the executive branch to coordinate science and technology policy across the diverse entities that make up the Federal research and development enterprise. Chaired by the President, the membership of the NSTC is made up of the Vice President, the Director of the Office of Science and Technology Policy, Cabinet Secretaries and Agency Heads with significant science and technology responsibilities, and other White House officials.

A primary objective of the NSTC is the establishment of clear national goals for Federal science and technology investments in a broad array of areas spanning virtually all the mission areas of the executive branch. The Council prepares research and development strategies that are coordinated across Federal agencies to form investment packages aimed at accomplishing multiple national goals. The work of the NSTC is organized under four primary committees: Science, Technology, Environment and Natural Resources and Homeland and National Security. Each of these committees oversees subcommittees and working groups focused on different aspects of science and technology and working to coordinate across the federal government.

For additional information concerning the work of the National Science and Technology Council please visit www.ostp.gov/cs/nstc.

About the Office of Science and Technology Policy

The Office of Science and Technology Policy advises the President on the effects of science and technology on domestic and international affairs. The office serves as a source of scientific and technological analysis and judgment for the President with respect to major policies, plans and programs of the Federal Government. OSTP leads an interagency effort to develop and implement sound science and technology policies and budgets. The office works with the private sector to ensure Federal investments in science and technology contribute to economic prosperity, environmental quality, and national security. For more information visit http://www.ostp.gov.

About the Office of Management and Budget

The predominant mission of the Office of Management and Budget (OMB) is to assist the President in overseeing the preparation of the Federal budget and to supervise its administration in Executive Branch agencies. In helping to formulate the President's spending plans, OMB evaluates the effectiveness of agency programs, policies, and procedures, assesses competing funding demands among agencies, and sets funding priorities. OMB ensures that agency reports, rules, testimony, and proposed legislation are consistent with the President's Budget and with Administration policies.

In addition, OMB oversees and coordinates the Administration's procurement, financial management, information and regulatory policies. In each of these areas, OMB's role is to help improve administrative management, to develop better performance measures and coordinating mechanisms, and to reduce any unnecessary burdens to the public.

For more information visit http://www.whitehouse.gov/omb/.

The Education Subcommittee

The Education Subcommittee (Ed Sc) of the National Science and Technology Council's Committee on Science advises and assists the Committee on Science and the NSTC on education policies, procedures and programs relating to Science, Technology, Engineering, and Mathematics (STEM) education and workforce development; the scientific research base and methodological approaches for evaluating and improving STEM education programs; and current, new and evolving strategies in public and private sectors for improving the teaching and learning of STEM education. The Subcommittee will address education and workforce policy issues and research and development efforts that focus on STEM education issues at the Pre-K-12, undergraduate, graduate, postdoctoral and lifelong learning levels, as well as current and projected STEM workforce needs, trends and issues.

About this document

This report details the progress made by Federal agencies toward implementing the recommendations of the Academic Competitiveness Council, specifically, that agencies with STEM education programs identify high leverage programs and collaborate on how to structure evaluations, embed metrics into their programs and coordinate their activities. In accordance with the recommendations of the May 2007 ACC report, the information contained in this document was presented at a principals meeting chaired by Secretary of Education Margaret Spellings on March 14, 2008. For information on the ACC report please visit: http://www.ed.gov/about/inits/ed/competitiveness/acc-mathscience/index.html.

Report prepared by
Education Subcommittee

Committee on Science
National Science and Technology Council (NSTC)

Members of the NSTC Education Subcommittee

Co-chairs
- Duane Alexander (NIH)*
- Cora Marrett (NSF)*
- Grover (Russ) Whitehurst (ED)*

Department of Agriculture
- George Cooper*

Department of Commerce
- Louisa Koch (NOAA)*
- Kim Benson (NOAA)
- Susan Heller Zeisler (NIST) ^
- Hratch Semerjian (NIST)*

Department of Defense
- Robert McGahern*^

Department of Education
- Grover Whitehurst*

Department of Energy
- William Valdez*
- Cindy White

Department of Health and Human Services
- Duane Alexander (NIH)*
- Bruce Fuchs (NIH)
- Annette Debisette*
- Joan Weiss (HRSA)^

Department of Homeland Security
- Desiree Linson^
- Larry Morgan

Department of Interior
- Robert Ridky (USGS)*

Department of Labor
- Katie Anderson^
- George Weltz*
- Jennifer McNelly

Department of State
- Andy Reynolds*

Department of Transportation
- Michael Avery^
- Kelly Leone*

Department of Veterans Affairs
- Malcom Cox*

Environmental Protection Agency
- Cynthia Nolt-Helms*
- Paul Juengst*^

Executive Office of the President
- Irma Arispe (OSTP)*
- John P. Bailey (DPC)*
- Irene Kariapuzha (OMB)*
- Kathy Stack (OMB)^

National Aeronautics and Space Administration
- James Stofan*^
- Joyce Winterton*

National Science Foundation
- Bernice Anderson^
- John (Spud) Bradley^
- Marta Cehelsky^
- Joan Ferrini-Mundy*
- Cora Marrett*
- Tia McNair^

Smithsonian Institution
- Stephanie Norby
- Sally Goetz Shuller*

Special Acknowledgements to additional contributors and participants in the Education Subcommittee

Marlene Kaplan (NOAA), Claire Saundry (NIST), Paul Jesukeiwicz (DOD), Jeffrey Johnson (ED), Jeffery Dilks (DOE) Dan Berch (NIH), Diana Espinoza (HRSA), Michelle Richardson (HRSA), Jennifer Sutton (NIH), Allison Cole (OMB), Diane DiEuliis (OSTP), Jennifer Gera (OMB), Amy Kaminski (OMB), David Orr (OMB), Malcom Phelps (NASA), Robert Shea (OMB)

Table of Contents

Finding Out What Works:
Agency Efforts to Strengthen the Evaluation of Federal STEM Education Programs

A Report of the National Science and Technology Council
Education Subcommittee

"And to keep America competitive, one commitment is necessary above all: We must continue to lead the world in human talent and creativity. Our greatest advantage in the world has always been our educated, hardworking, ambitious people -- and we're going to keep that edge. Tonight I announce an American Competitiveness Initiative, to encourage innovation throughout our economy, and to give our nation's children a firm grounding in math and science."

President George W. Bush
State of the Union Address
January 31, 2006

I. Introduction

On February 8, 2006, President Bush signed into law the Deficit Reduction Act of 2005 (P.L. 109-171). Section 8003 of the Act, under Section 401A(a)(2) of the Higher Education Act of 1965, established the Academic Competitiveness Council (ACC).

The statute mandated that the Secretary of Education chair the Council and that its membership consist of officials from Federal agencies with responsibility for managing Federal mathematics and science education programs. The law charged the ACC with the following tasks:

- Identify all Federal programs with a mathematics or science education focus;
- Determine the effectiveness of those programs;
- Detect areas of overlap or duplication among those programs;
- Recommend ways to efficiently integrate and coordinate those programs.

In May of 2007 the ACC released its final report and made six recommendations to improve, integrate and coordinate Federal STEM education programs (see Appendix A) .[1] The final recommendation was that agencies with science, technology, engineering, and mathematics (STEM) education programs collaborate to implement ACC recommendations under the auspices of the National Science and Technology Council (NSTC) of the Office on Science and Technology Policy (OSTP). The NSTC Committee on Science (COS) formed an Education Subcommittee (Ed Sc) to carry out this task. The subcommittee is co-chaired by Dr. Russ Whitehurst from the Department of Education (ED), Dr. Cora Marrett from the National Science Foundation (NSF), and Dr. Duane Alexander from the National Institutes of Health (NIH).

This report details the progress made by Federal agencies toward implementing the recommendations of the Academic Competitiveness Council. Federal agency activities described here span the ACC recommendations; however, this report pertains most directly to recommendation VI that agencies with STEM education programs should identify high leverage programs and collaborate on how to structure evaluations, embed metrics into their programs and coordinate their activities.

[1] U.S. Department of Education. Report of the Academic Competitiveness Council. May 2007.

In accordance with the recommendations of the May 2007 ACC report, the information contained in this document was presented at an ACC principals meeting chaired by Secretary of Education Margaret Spellings on March 14, 2008.

II. Approach and Methodology

Following issuance of the ACC report, consultation began between OSTP, the Office of Management and Budget (OMB) and three co-chairing agencies (Department of Education, NSF, and NIH). This resulted in reconstitution of an Education Subcommittee of NSTC and a plan for collection of data from federal agencies for the development of a report to the President and the ACC on progress in meeting ACC recommendations.

Development of a data call: Concurrent with the efforts to reconstitute the Ed Sc, OSTP, OMB and the three co-chairs developed a draft evaluation template to capture information on agency implementation activities for the ACC recommendations (see Appendix B). The template builds directly on the work of the ACC by integrating the concept of a high-leverage program, the list of goals and metrics that could be used by the agencies to evaluate program effectiveness, and outlining a hierarchy of designs.

The template called for each agency participating in the ACC to provide 1) Information on at least one "high leverage program" and (2) General/aggregate information on evaluation activities for the agencies' overall portfolio. A high leverage program is defined as one with significant potential to enhance student learning, strengthen teacher quality, increase the number of postsecondary students who complete STEM degree programs, and/or programs that add substantially to the knowledge base of effective innovation practices in STEM education. In addition, it was noted that programs proposed for an increase of more than $10 million in the President's Budget are assumed to be high leverage. Agencies with more than one such program were asked to submit multiple templates or an explanation for why the program proposed for expansion is not considered high leverage. The template collected information in five categories:

- General Information: Program name, description and discussion of why the program is high leverage.
- Outcome Measures: national and common metrics (Agencies were encouraged use metrics included in the ACC report and, if others were used, to describe and justify them.)
- Description of How Measures are Used in Program Operations: frequency of reporting, methods to ensure quality, etc.
- Evaluation: information on whether the program has been evaluated in the past, plans for future evaluation, key research questions, which tier of the ACC-specified hierarchy of study designs the planned evaluation falls under (experimental, quasi-experimental, or other designs), and methods for ensuring the quality of the design, including securing external evaluators with appropriate expertise.
- Disseminating and Using Evaluation Results: including efforts to highlight effective or ineffective practice
- Overall Agency Progress on Implementing ACC Recommendations: information about how agencies are implementing evaluation for the portfolio as a whole (i.e., beyond the high-leverage program).

To assist agencies in completing the evaluation template and to facilitate the collection of data in a uniform format, the Institute of Education Sciences within the Department of Education developed a model response that was circulated along with the blank template.

Establishment of the NSTC Education Subcommittee: The Ed Sc is comprised of representatives from agencies within the NSTC Committee on Science, and as indicated previously, chaired by Education, NSF and NIH. Agency representatives include individuals with (1) A substantive knowledge of STEM education programs within their agency's portfolio, and (2) Experience with evaluation research and/or the development and application of performance measures. To ensure appropriate expertise, agencies were allowed to nominate more than one representative to the subcommittee and, in doing so, to designate a key or primary representative. As with all NSTC Subcommittees, OSTP and OMB are active participants in the work of this group.

Responding to the ACC recommendation is a critical initial task of the Ed Sc, but it is expected that this subcommittee will address a broad range of issues related to STEM education, including efforts by working groups in the areas of education research, graduate and postgraduate education, and human resources development, among others. In addition to the ACC-related activity described above, the subcommittee will provide a forum for the exchange of information and expertise about the integration of rigorous evaluations in STEM education program design and assessment. The subcommittee will also monitor progress of the collective Federal effort in meeting the national and program goals identified by the ACC, and will facilitate improved coordination among agencies that sponsor STEM education programs targeted toward similar goals or that serve similar populations.

During the first subcommittee meeting, held on June 27, 2007 Dr. Sharon Hays, Associate Director for Science, OSTP, explained the background for reconstituting the subcommittee and the subcommittee's charge. Robert Shea, Associate Director for OMB Management and Government Performance, outlined the ACC recommendations and provided an overview of the draft evaluation template. Agencies were asked to provide comments or suggest changes to the template prior to its distribution.

The data call was released on July 2, 2007. Although both ACC and non-ACC agencies comprise the Ed Sc, only ACC agencies were asked to complete the template. A web page was set up on the www.max.omb.gov site to house agency responses.

The Evaluation Subgroup: A subgroup of Ed Sc members participated in the discussion of evaluation templates. The Evaluation Subgroup was led by Russ Whitehurst (ED), Bruce Fuchs (NIH), and Joan Ferrini-Mundy (NSF). The purpose of the group was to facilitate collegial discussions leading to the improvement of proposed evaluations.

The Evaluation Subgroup met on July 25, August 13, August 22, September 5, and October 12 to discuss agency submissions. Each agency representative described their agency's high leverage program and evaluation plan. Members of the evaluation subgroup asked questions and provided suggestions for improving the evaluation design, after which agency representatives revised their plans and resubmitted them based on this feedback. The evaluation subgroup was not responsible for approving the resubmissions[2]. The data collection period ended on October 5, 2007.

[2] Note: due to difficulties in scheduling, the DOD template was not discussed.

The process used by the Evaluation Subgroup to respond to the ACC recommendation prompted increased awareness within agencies of the need for more rigorous evaluation efforts and stimulated the initiation of efforts to meet this need. The process also revealed that some agencies lack the human capacity to design and implement rigorous evaluation efforts. This is not surprising given the uneven nature of evaluation planning and execution prior to the work of the ACC. (It should be noted that these capacity issues are not unique to federal agencies and have been discussed across the entire education enterprise.) With the heightened awareness of the importance of evaluation it is important that agencies have, or develop, the capacity to plan and implement rigorous evaluations of their education programs.

The remaining sections of this report summarize themes across the agency evaluation plans and recommendations stemming from this effort.

III. Agency Actions to Strengthen Evaluation Rigor in High Leverage Programs

One of the goals of the Education Subcommittee is to encourage scientifically rigorous evaluations of STEM education programs in order to advance evidence-based policy and practices. The Subcommittee shares the concerns regarding the STEM workforce in the federal government and, more generally, the STEM skills of the U.S. workforce that have recently commanded much public attention. Successful, large-scale efforts to improve STEM education should be based on the existing body of knowledge generated from longitudinal educational research studies. Federal agencies will want to conduct both program-level and project-level grantee evaluations more rigorously.

A large portion of the Education Subcommittee's time was spent discussing ways to improve the evaluation of agencies' STEM education programs. The task was made more complex by the fact that many agencies' STEM education programs have unique educational goals related to their specific scientific mission. Also not all participating agencies are at the same stage with respect to the maturity of their STEM education programs, or their plans for evaluating those programs. However, this time was well spent, as it resulted in a new level of understanding and agreement across the agencies. ***(See Appendix C for a more detailed treatment of some of the issues that were discussed.)***

It is important to choose evaluation methods that are appropriate to each agency program's stage of development and to the research questions being asked about a particular educational intervention. The Education Subcommittee is not prescribing a "one-size fits all" approach. Scientifically valid education evaluations can employ a range of methodologies. Ultimately, however, an educational intervention will reach a stage of development and maturity at which its effectiveness should be assessed; i.e., is the intervention "working"? At this stage, the randomized controlled trial (RCT), when appropriately and correctly implemented, is the most powerful design for detecting these effects. However, because RCTs are not always feasible, other methods within a family of quasi-experimental designs can be utilized to estimate an intervention's impacts. It is also likely that additional evaluation techniques will be used in combination with these methods, as it is also important to understand *why* a particular intervention does or does not have an effect.

While the Education Subcommittee wants to encourage increased use of RCTs , and methodologically strong quasi-experiments, policy makers should also understand the ultimate goal. Establishing causality in education science requires a coherent body of theory that can be used to predict specific relationships between program interventions and student outcomes. A single small RCT, no matter how well designed, will not be sufficient to establish general

conclusions that a program or intervention works. The ability to generalize findings requires a large body of evidence gathered in different settings and circumstances. Expert review of the accumulated research, and careful consideration of the validity and relevant characteristics of each study to substantiate the results and outcome, will also be a necessary step. The Education Subcommittee is committed to helping the agencies achieve this goal.

What are models used within the federal government's high leverage STEM programs?

A total of 17 high leverage programs were identified which varied considerably in size ranging from small pilot efforts such as the NIST-MCPS program ($100,000) targeted to one school district (Montgomery County, Maryland) to the ED's Math Now program ($250,000,000). Appendix C lists the programs and key descriptive information. Detailed agency responses may be found in Appendix D.

Table 1
High Leverage STEM Education Programs

Agency/Program	Target	FY07 Enacted	FY08 Enacted	FY09 Budget
Education				
MATH NOW	K-12	$0	$250,000,000	$95,000,000
SMART	Undergraduate	*	*	*
AP/IB	K-12	$37,026,000	$122,175,000	$70,000,000
ATP	K-12	$0	$25,000,000	$10,000,000
NIH				
Science Education Partnership Award (SEPA)	K-12, informal education and outreach	$16,009,000	$15,325,000	$16,009,000
NSF				
Discovery Research K-12 (DR-K12)	K-12	$97,000,000	$107,000,000	$108,000,000
DOE				
Science Undergraduate Laboratory Internship (SULI)	Undergraduate	$2,719,000	$2,876,000	$2,600,000
NOAA				
The JASON Project	K-12	$1,900,000	$1,000,000	$0

NIST				
NIST-MCPS Pilot Summer Institute	K-12	$100,000	$114,000	
NASA				
NASA Explorer Schools	K-12	$8,700,000	$12,200,000	
DOT				
University Transportation Centers (UTC)	Undergraduate, graduate, postgraduate	$67,030,000	$76,700,000	$77,000,000
SI				
Science and Technology Concepts Program (STC-elementary and middle school)	K-12	$882,903	NA	
HRSA				
Nursing Workforce Diversity Program (NWDP)	K-12, undergraduate, CNAs, LPNs and adults from disadvantaged backgrounds interested in pursuing a nursing degree	$16,100,000	$16,100,000	
EPA				
P3 Program	Undergraduate and graduate	$1,200,000	$1,200,000	
DHS				
Scholarship and Fellowship Program	Undergraduate, Graduate, Postgraduate	$13,000,000	$7,000,000	$7,000,000
DOD				
Pre-Engineering Partnerships	K-12	$0	$13,000,000	

The high leverage programs also varied considerably in approach. For example, while most (12 of the 17) targeted K-12, these programs used varied educational approaches including formal education, informal education, or multiple approaches. High leverage programs are grouped into the six general descriptive categories seen below. These categories are provided for general descriptive purposes only. These descriptive categories are not mutually exclusive and some programs fit in more than one category. Similarly, some programs are not perfect matches with the general categories in which they appear.

1. Grants to state or local educational agencies (SEAs and LEAs): These grants support such things as instructional programs, professional development, and the development of assessments. Included in this category are the following. The Department of Education's MATH Now as authorized under the COMPETES Act would provide competitive grants to SEAs to fund LEAs to improve mathematics achievement of elementary and middle school students. The Department of Education's Adjunct Teacher Program makes competitive grants to partnerships of school districts and states, or appropriate public or private sector institutions, to create opportunities for professionals with STEM disciplinary expertise to teach secondary school courses in math, science, or a critical foreign language. The Advanced Placement/International Baccalaureate (AP/IB) program will provide grants on a competitive basis to SEAs, LEAs, or partnerships of SEAs or LEAs with nonprofit organizations to fund teacher professional development, course development, and other activities designed to increase the number of students in high need schools who enroll in AP/IB courses in mathematics, science, or critical foreign languages. Other grant programs in this category are more formula driven. Examples are: the Department of Education's Science and Mathematics Access to Retain Talent (SMART) program; and the Math and Science Partnerships (MSP) program. The SMART program provides stipends to undergraduates who are eligible for a Federal Pell Grant and who are majoring in physical, life, or computer sciences, mathematics, technology, engineering, or a foreign language determined to be critical to national security. The MSP program supports state and local efforts to improve elementary and secondary students' achievement by promoting strong teaching skills. It is important to note that the latter are more difficult to evaluate in part because of the difficulty in establishing control groups. These programs are the most directly linked to improving student performance on NAEP and state assessments.

2. Agency partnerships to increase student and/or teacher interest and expertise in STEM disciplines: These programs focus on bringing together students and/or teachers with STEM professionals who act as teachers and mentors to increase interest and competency among the target population. These programs include the Department of Energy's Science Undergraduate Laboratory Internship program, which provides a diverse group of approximately 340 undergraduate students with an individually mentored research experience at one of the Department of Energy's National Laboratories as a way to increase preparedness for the STEM workforce. Also in this category is the National Institute of Science and Technology's Montgomery County Public Schools Summer Institute program which brings teachers together with NIST scientists to experience measurement research in an applied, or real world, setting.

3. Agency partnerships for professional development and/or implementation of curricula: Like the programs in category two, these high leverage programs encourage interest and engagement in STEM science, however, they also involve curricular

development and/or implementation. Examples include NOAA's JASON project, which seeks to increase middle level learners' science proficiency and inspire and motivate them to make science part of their education and career plans, and the Smithsonian Institution's National Science Resources Center (NSRC), whose efforts to improve science learning and teaching in the United States include the development of instructional materials and teacher professional development. Other examples include the NASA Explorer Schools (NES) program and DOD's Pre-engineering Partnerships (PEP) program NES establishes three-year partnerships between NASA and school teams, consisting of teachers and education administrators from diverse communities across the country. The NES includes professional development activities and provides educators the sustaining support. The PEP program supports partnerships among teachers, school districts, Service and DoD laboratory scientists & engineers, and K-12 teacher training universities to enhance the learning experience for middle school level students, grades 6-9.

4. Grants for research and development: These grant programs are related to student and teacher learning, to learning resources and models for students and teachers, and competitive design efforts. The NSF Discovery Research K-12 program is the largest program of this type, sponsoring research about and/or development of innovative resources models and technologies for use by students, teachers and policy makers. Another example is EPA's People, Prosperity and the Planet (P3) program, a two-phase grant competition program for institutions of higher education. Recipients use the money to research and develop their design projects during the academic year. Then all P3 grant recipients attend the National Sustainable Design Expo featuring the EPA's P3 Award competition on the National Mall in Washington, D.C. Although primarily a research program, P3 also includes requirements for integration of sustainability concepts as an educational tool and reporting on the quantifiable benefits to people and the planet.

5. Direct funding to students or to institutions for completion of education: This category of projects aims directly at the STEM pipeline by funding students to pursue STEM careers. They include the Department of Education's SMART grants, which are described in Category 1, HRSA's Nursing Workforce Development Program (NWD), and the Department of Homeland Security's Scholarship and Fellowship (S&F) Program. The NWD program provides grant support to increase nursing education opportunities for individuals from disadvantaged backgrounds through retention activities, pre-entry preparation strategies, and by providing student scholarships or stipends. The S&F Program provides scholarships for undergraduate and fellowships for graduate students pursuing degrees in DHS mission-relevant fields.

6. Grants to increase career opportunities: Two very distinct programs are included in this category - The DOT's University Transportation Centers (UTC) program and NIH's Science Education Partnership Award (SEPA) program. The UTC Program supports university-based centers of excellence to advance U.S. technology and expertise in the many disciplines comprising transportation through education, research and technology transfer. The educational activities relate to transportation and include multidisciplinary course work and participation in research. SEPA is a grant program that provides 5-years of funding for K-12 educational programs designed to increase career opportunities in science for children and to deliver topical and interactive information about NIH-funded medical research and an understanding about healthy living habits to the general public.

How are agencies improving the rigor of program evaluation since the issuance of the ACC report?

A range of programs, a range of designs: The diversity of objectives, structures, approaches, and target audiences among Federal high leverage programs calls for a range of evaluation approaches. Most programs incorporate some element of experimental or quasi-experimental design and many of the larger or more complex/diverse programs incorporate multiple evaluation design components. Four programs (Math Now, SEPA, DR-K12, and STCP) will employ Tier I experimental designs for at least one component of the overall program. In the case of Math Now, an RCT design is planned for the national program. SEPA and DR-K12, both grant programs, have incorporated the requirement for evaluation into solicitations and both programs will be enhancing those requirements to align with the ACC report's emphasis on increased rigor.

Programs such as Department of Education's Math NOW program, the NSF's DR-K12, NIH's SEPA, and the Department of Energy's SULI are striving to incorporate both project and program level rigorous evaluation using experimental or quasi-experimental designs and include additional data obtained from national surveys or other qualitative information.

In total, 12 of the 17 programs use or plan to use either Tier I (randomized control trials) and/or Tier II (quasi-experimental designs) in the coming years. Some high leverage programs, such as NASA's NES, NIST-MCPS, SI's STCP, and NOAA's JASON have been in existence prior to the ACC, and although evaluative information was collected in the past, the programs did not incorporate rigorous evaluation designs. These agencies reported that they are now moving toward more rigorous Tier II designs.

Two new programs are under development (ED's ATP and DOD's PEP) and although they have no formal evaluation designs at present, a percentage of program resources have been committed to evaluation.

Some agencies reported the incorporation of other evaluative approaches. For example, three programs (SEPA, DR-K12, and DOT's UTC) support the development of research, innovative resources or tools, and educational practices. The output of such endeavors may take the form of a curriculum or the publication of research findings. To go beyond output measures (for example, the number of participants or products), these programs are using an expert review approach to evaluate the results of research endeavors. The purpose of the expert review is to provide an independent assessment of the technical and scientific merit of the research. Further, the SEPA program is utilizing a Peer Evaluation Cluster (PEC), pioneered by the Howard Hughes Medical Institute (HHMI). In this model, four SEPA project awardees form a PEC that will conduct sequential evaluation site visits on each individual member. The process of being an evaluator for 3 visits and being evaluated on one visit allows the participants to develop expertise in how to structure evaluations, to fine tune the metrics used and to share best practices. Many of the SEPA awardees also have HHMI funding and, having participated in the PEC program, are qualified to instruct the SEPA community on implementation of the PEC process. This process will complement SEPA programmatic changes to increase evaluation rigor and coordinate evaluation metrics across the SEPA program.

Linking program goals to evaluation questions: An essential first step in integrating evaluation into program design is developing evaluation questions that are linked to program goals. Agencies were asked to identify the key research questions that would be addressed in

their evaluation projects. Within the 17 high leverage programs, research questions centered on

1) The direct effects of educational programs and interventions on student achievement, enrollment in and/or graduation from universities with a STEM major, on preparation for work in STEM fields, recruitment into and retention in the STEM work force.

2) Other important measures that are less directly linked to academic performance: the effect on teacher development and subsequently on student performance, student engagement, student knowledge about STEM careers.

3) Innovations in the form of models, resources, technologies, and innovative research completed by those in STEM research programs.

Identifying appropriate metrics: A key contribution of the ACC report was the identification of goals and metrics. The varied structure of the 17 high leverage programs described here influences the type of metric used (student, project, program, or a combination of metrics). The metrics for individual programs are described in Appendix D, the Agency Evaluation Templates; however, metrics reported in the data call may be grouped broadly into the following categories:

Table 2
Categorical summary of metrics used or proposed by agencies
For their high leverage program

Metric type	Programs using related metrics
Improving student performance (improving test scores, pass rates, achieving specified performance levels)	9 programs: Math Now, AP/IB, ATP, SEPA, DR-K12, SULI, JASON, STCP, SWDP, DOD.
Increasing the number of students entering STEM fields (increasing enrollment, graduation, majors; increasing the number of individuals in STEM graduate training; increasing the number of graduates who take jobs in STEM fields)	6 programs: SMART, AP/IB, SEPA, SULI, DOT, DHS S&F
Increasing teacher competency: (increasing the number of qualified teachers, increasing the number with STEM undergraduate or graduate training)	7 programs: AP/IB, ATP, SEPA, DR-K12, JASON, NIST-MCPS, STCP
Increasing student participation in sustained extracurricular activities.	5 programs: SEPA, JASON, NIST-MCPS, NES, NWDP
Increasing student interest, enjoyment in STEM education fields	6 programs: SEPA, JASON, NIST-MCPS, NES, STCP, P3
Effective approaches for learning: (Number of funded new approaches found to be effective; number of courses added; percentage of institutions incorporating sustainability into engineering curricula)	5 programs: DR-K12, ATP, P3, JASON, STCP
Expert evaluation of the products of research (papers, research findings, etc.)	3 programs: SEPA, DR-K12, UTC
Increased public awareness, attitudes	2 programs: SEPA, NES

Employer satisfaction with STEM graduate	1 program: SULI

The metrics selected are a function of the program purpose and design, with those most directly targeted at K-12 education and most aligned with state standards applying measures that track improvements in NAEP and state assessments. When attitudinal measures (such as student enjoyment) are used by programs, they are used in conjunction with other measures and in the context of a broader evaluation. Programs producing "products" (e.g., models, tools, innovative approaches, research findings) are incorporating evaluative aspects such as expert review of products for technical and scientific quality (SEPA and DOT), and inclusion of evaluative testing to ensure that programs are good candidates for scale up (DR-K-12).

Development of comparison groups: Several programs did not initially envision comparison groups; however, as a result of the dialogue of the Evaluation subgroup, the programs are modifying designs to include such a group. The most frequent scenario for this case is the use of a "wait" group: a competitively selected group is funded or receives or a specific intervention , while a second group is delayed for a year. The second, or wait, group serves as the comparison. One project in particular, the NES program, made use of a natural experiment in which insufficient first year funding inadvertently caused creation of a "wait group".

Options to advance rigorous evaluation: The ACC report identified five options to advance rigorous evaluation: Competitive Priority, Required of All Applicants, Cross Project Evaluation, Sheltered Competition, and Waivers to Allow Impact Study[3]. The most frequently used approach is Cross Project Evaluation, used or planned for use by four programs (Math Now, AP/IB, SULI, UTC and P3). Competitive Priority is the next most frequently used (for Math Now, SEPA, and STCP). One project, JASON, proposes Sheltered Competition, and models have not yet been established for the remaining projects.

Planning for effective dissemination and use of evaluative information: The most commonly expressed vehicle for disseminating information from an evaluation is through web posting. Seven programs (Math Now, AP/IB, SEPA, NES, UTC, P3, STCP) are either using or

---

3 ***Options to advance rigorous evaluation: Definitions.***

Competitive Priority*. The program gives priority consideration to award applicants that propose to conduct a scientifically-rigorous evaluation of their project. Such applicants are given additional points in the proposal evaluation process, and may also be awarded additional funds to conduct the evaluation.*

Required of all applicants*. The program requires award applicants to conduct a scientifically-rigorous evaluation of their project, and awards them additional funds to conduct the evaluation. Agency issues standards to govern quality of evaluations.*

Cross-project evaluation sponsored by the program*. The program or agency itself sponsors a scientifically-rigorous evaluation of one or more distinct interventions (e.g., a specific course curriculum) that a number of program awardees have adopted. The program or agency selects an independent researcher team to conduct this cross-project evaluation. The program requires its awardees to participate in such evaluations if asked.*

Sheltered competition*. The program sets aside a portion of its funds to conduct a "sheltered competition" for funding awards to implement a specific intervention that the program seeks to evaluate (e.g., a well-defined teacher training model that a federal teacher professional development program seeks to evaluate). The program then selects an independent research team to conduct a scientifically-rigorous evaluation of the intervention, and requires the selected awardees to participate in the evaluation.*

Waivers to allow impact study*. The agency or program waives provisions of law or regulation to allow program awardees to carry out demonstration projects of new interventions (e.g., new methods of program delivery), and in return requires such awardees to conduct a scientifically-rigorous evaluation of their demonstration project. (This option is more applicable to formula grant rather than discretionary grant programs.)*

proposing to use the web for dissemination. Four agencies proposed specific conferences for grantees (AP/IB, SEPA, NSF, EPA). Grant programs propose using evaluative information for improving future solicitations (Math Now, NSF), while programs whose purpose is to develop curricula will use evaluative information for improvement of curricula (JASON, STCP)

What evaluation resources would agencies like to have had? The Evaluation Subgroup discussions identified areas where future collaborations would be helpful and where evaluative resources are needed. Desired resources include: (1) a method (for example, a long-term data base) to track students across their educational trajectory and into the workforce to evaluate the effect of STEM training; (2) information on the implementation of interventions, so that agencies can not only identify effective approaches but learn about the best ways to ensure their successful application; (3) common or shared metrics so that similar programs, or program components, can be benchmarked against each other; (4) the ability to align curriculum with actual test items (rather than standards or test items) so that the effects of educational programs can be directly measured; and (5) technical and financial resources to plan and carry out evaluations.

IV. Agency Progress: Summary of Actions To Implement ACC Recommendations

Agency submissions indicate that work is underway across the Federal government to align program outcome expectations with the metrics developed by the ACC, to increase the quality of evaluation efforts, and to collaborate with other agencies, both Federal and non-Federal, on evaluation. Progress varies considerably from agency to agency and the extent of progress seems to be related to the extent of involvement of the agencies in the work of the ACC. Those agencies that most actively participated in the ACC seem to be further along in aligning their evaluation efforts with the recommendations of the ACC.

Agencies continue to align their evaluations with the ACC metrics and are developing more rigorous evaluations. Most are convening workshops or establishing special units or groups to move this work forward. Many are engaging nationally recognized evaluation experts in their efforts. Other agencies are relying on existing parts of the organization and revisiting methods that were already in place. Of the agencies actively working through workshops or special groups, some are further along than others. Again, the extent of progress seems to be related to the level of participation in the ACC and possibly with the capacity and past experience within the agency to deal with evaluation issues.

In spite of the fact that progress is not uniform, it is evident that Federal agencies involved in STEM education programs have taken this work seriously and are moving forward. It is likely, however, that some agencies will need to devote more resources to evaluation and that they will need to work collaboratively with those that are further along and that have more experience and capacity in this area. In many cases additional funding will need to be identified to support more ambitious evaluation efforts.

Table 3
A summary of agency responses on actions to implement ACC recommendations

Agency	I. Common Metrics	II. Evidence-based focus	III. Coordination
Department of Defense	The assessment team is integrating	A plan for increasing rigor in	Successful approaches used

	ACC metrics into program evaluation.	evaluations is being implemented.	previously are being implemented.
Department of Education	Performance measures have been developed for currently funded programs and potential measures are being developed for new programs. These are aligned with the ACC metrics.	Efforts are under way to improve the quality of evaluations, especially for smaller programs. Also, a RCT is being conducted of commonly used mathematics texts.	Data on project effectiveness is made available to staff and the public. A reporting tool is being developed that could be used by other agencies.
Department of Energy	Development of metrics is underway.	Rigorous evaluation framework has been developed and reviewed by external peer panel.	Planning is part of a larger strategic plan that has been developed, along with detailed program implementation plans.
Department of Health and Human Services	Two existing methods of evaluation are used. Outcome data are entered into a common system that allows aggregation.	Not addressed	Not addressed
Department of Homeland Security	Data are collected on the number of students in the Fellowship program.	A program evaluation is planned for FY 2008.	N/A
Department of Transportation	The Director of the UTC program oversees evaluation, including use of ACC metrics.	A contract is to be let to conduct rigorous evaluations.	Coordination will be done through conferences & posting evaluation results on the web.
Environmental Protection Agency	The Board of Scientific Councilors (BOSC) is developing metrics.	The BOSC is developing an evaluation plan.	Not addressed
National Aeronautics & Space Administration	A budget line dedicated to evaluation has been established. Metrics were developed as part of	A contract will be let to conduct rigorous evaluations.	The evaluation manager will ensure coordination.

	NASA's participation in the ACC.		
National Institutes of Health	A working group will align metrics with ACC goals	The working group will work with program directors to strengthen rigor of evaluations.	Another working group will develop ways to improve coordination
National Institute for Science and Technology	Metrics will be aligned with ACC goals.	A plan for increased rigor will be implemented as resources allow and as more data for comparisons becomes available	Continued work within the ACC will be used to improve coordination.
National Science Foundation	Program and project metrics are being designed based on the ACC metrics.	The Directorate for Education and Human Resources is taking the lead in work to strengthen evaluation design across the agency.	Conferences are planned that will form the basis for ongoing conversations on evaluation.
National Oceanic and Atmospheric Agency	Participation in the ACC fostered a process to outcome measures consistent with the ACC metrics.	An evaluation plan is being developed that is consistent with the ACC recommendations on rigor.	Contacts developed through the ACC are being used to continue collaboration.
Smithsonian Institution, National Science Resources Center	For each of several types of programs, the Smithsonian has defined data collection methods and a system of reporting. An analysis is underway to align outcomes with ACC metrics.	Studies have been commissioned to develop rigorous evaluation methods.	An agreement has been signed with the Council of Chief State School Officers, with participation with the Dept. of Education, to foster appropriate collaboration.

V. Recommendations

Based on the agencies' discussions about high leverage programs and the implementation of more rigorous evaluations within their large STEM education portfolios, the following steps are

recommended as offering potential for the greatest impact on improving effectiveness and coordination of STEM education programs.

1. **To reap the benefits of the collaborative work conducted in 2007, NSTC agencies should move forward with implementation of their evaluation plans for high leverage programs in 2008 incorporating, where feasible, the recommendations for improvement made by the subcommittee.**

These evaluations will advance the Federal government's knowledge about the effectiveness of agency program investments and can inform future decisions by program managers and policymakers.

2. **To strengthen agency capacity to plan and carry out rigorous evaluations of their education programs, NSTC agencies should pursue cost-effective approaches to rigorous evaluation and work with OMB during the annual budget process to assess resource requirements.**

Most agencies do not currently have the in-house capacity to conduct the rigorous evaluations required. For some agencies this would require adding staff and/or hiring a contractor to provide advice and expertise. The establishment of a dedicated unit to provide advice and support for evaluation efforts would be useful.

3. **To facilitate interagency coordination and a shared focus on improving STEM education outcomes, the NSTC Education Subcommittee should maintain, update, and enhance the program database and ACC metrics to ensure its continued value and relevance to agencies, Congress, and outside organizations.**

The current ACC program database primarily contains higher level programmatic information such as program name, a brief description, budget information and contact information for the program director. While this information constitutes a useful beginning, currently information does not exist in sufficient detail to allow program directors with substantially overlapping interests to locate one another across agencies. A greater level of detail regarding programmatic details could foster interagency collaborations and promote efficiencies.

Achieving this vision will require a much more detailed database of program target audiences, goals, approaches, and methodologies employed. The development and maintenance of such databases is costly and complex, and this area will be a topic of attention for the NSTC Education Subcommittee.

4. **To continue constructive discourse on how best to assess the impact of STEM education programs, the NSTC Education Subcommittee should develop a clear and coordinated message on the role of evaluation and evidence-based research in strengthening STEM education that builds on the recommendations of the May 2007 ACC report.**

One of the goals of the Education Subcommittee will be to encourage improved evaluations of new and existing programs. The Subcommittee can accomplish this by promoting discussions across agencies about the best ways to interpret and apply the Subcommittee's definition of "rigorous evaluation" to their programs. The subcommittee may decide to sponsor

special workshops to help agency grantees understand the new emphasis on scientific program evaluation and learn how to apply these methods to their own programs.

5. **To fully develop promising evaluation models for similar programs intended to achieve similar outcomes, and to enable promising interventions to be rigorously tested, the NSTC Education Subcommittee should facilitate linkages across programs and agencies.**

The ACC process will eventually enable Federal program directors with similar interests to find one another, a process that will be accelerated by the creation of the database discussed in the third recommendation. The NSTC Education Subcommittee should serve as a platform for programs with mutual interests that have already been identified to work together. For example, a number of Federal agencies that conduct programs which place STEM teachers into scientific research laboratories for a summer work experience have been identified through the ACC process. The NSTC Education Subcommittee should sponsor a workshop that will make it possible to design an evaluation study that spans a number of agency programs to increase its power and sensitivity.

Interagency linkages between agencies may be needed to ensure that some promising interventions are rigorously tested. Some Federal programs support the initial development of new educational approaches and, for these, experimental or quasi-experimental evaluation designs may be premature. When such interventions show promise of having measureable impact there may be other Federal programs or agencies which can further develop them and subject them to rigorous evaluation to assess whether they should be scaled up and broadly disseminated.

6. **To strengthen agency capacity to assess long-term educational and workforce outcomes for postsecondary STEM programs, the NSTC Education Subcommittee should engage with other efforts already underway to foster appropriate consistency in the administration and evaluation of these programs.**

Many agencies are unable to assess the long term impact of their fellowship and postdoctoral programs because of inadequate longitudinal data on whether STEM graduates are entering and remaining in STEM fields. Having each agency create its own process for collecting and analyzing longitudinal data would be inefficient and cost-prohibitive. There is an opportunity to address this challenge in the coming year by synchronizing the Education Subcommittee's work with ongoing interagency efforts that are modifying fellowship applications and reporting forms (e.g., as part of the Grants.gov initiative) and improving the collection and utilization of data on postsecondary STEM education and outcomes.

VI. Conclusion

The process begun by the Deficit Reduction Act of 2006 that created the Academic Competitiveness Council (ACC) and the May 2007 ACC Report has led to a rapid transformation of Federal agency actions and attitudes toward how to manage and evaluate STEM education programs. Federal agencies are now focused on rigorous program management and evaluation that should pay dividends in the long-term to U.S. taxpayers and program participants.

This process, however, has just begun and is now being carried forward by the NSTC Subcommittee on Education. The immediate concerns of developing a common understanding of what it means to evaluate programmatic success and what that success should look like have largely been addressed. The NSTC Subcommittee on Education will now turn its attention to implementing four key recommendations contained it this report that will institutionalize the gains that have been made to date. Those recommendations focus on providing Federal agencies with the data, resources, and linkages that are required to ensure that appropriate evaluation and program design are built into all Federal STEM education efforts in the future.

Appendix A
Recommendations of the ACC

Recommendation 1: The ACC program inventory, goals and metrics should be living resources, updated regularly and used to facilitate stronger interagency coordination.

Recommendation 2: Agencies and the federal government at large should foster knowledge of effective practices through improved evaluation and-or implementation of proven-effective, research-based instructional materials and methods.

To improve outcomes, agencies will focus their attention on:

- Measuring the impact of STEM education programs using the ACC goals and metrics;
- Implementing more rigorous evaluations, consistent with the hierarchy of evaluation designs presented in this report, to assess whether programs or activities are having the intended, positive impact;
- Implementing proven practices that have shown success through scientifically evaluated evidence; and
- Disseminating widely, within the federal government and to the public, consistent information on the effectiveness of federal programs.

Recommendation 3: Federal agencies should improve the coordination of their K–12 STEM education programs with states and local school systems.

Recommendation 4: Federal agencies should adjust program designs and operations so that programs can be assessed and measurable results can be achieved, consistent with STEM education program goals.

Recommendation 5: Funding for federal STEM education programs designed to improve STEM education outcomes should not increase unless a plan for rigorous, independent evaluation is in place, appropriate to the types of activities funded.

Recommendation 6: Agencies with STEM education programs should collaborate on implementation of ACC recommendations under the auspices of the NSTC. Specifically, NSTC member agencies should identify high-leverage programs and collaborate on how to structure evaluations, embed metrics into their programs, and coordinate their activities. Under the auspices of the NSTC, member agencies will present a report to the President on agency progress and additional detailed recommendations at an ACC principals meeting chaired by the Secretary of Education by Oct. 1, 2007.

Appendix B
Agency Evaluation Template used for July 2007 Data Call

Agency Name:

I. GENERAL INFORMATION.

Program Name:

2007 Funding:

2008 President's Budget:

Primary program subgroup: ***(choose from K-12, Undergraduate, Graduate/Postgraduate, or Informal Education and Outreach)***

Program description: ***(use or update description in ACC inventory)***

High-leverage program: (*Briefly explain why this is considered a high-leverage program with significant potential to enhance student learning, strengthen teacher quality, increase the number of postsecondary students who complete STEM degree programs, or add substantially to the knowledge base about effective innovation in STEM education. [Note: Programs proposed for an increase of more than $10 million in the President's Budget are presumed to be high leverage. If an agency has more than one such program, it should submit multiple templates or an explanation for why the program to be expanded is not considered high leverage.]*

II. OUTCOME MEASURES.

National metric: (*Identify the primary national metric that corresponds to this program. If applicable, select other national metrics. If the metric is not from Appendix B of the ACC report, describe and justify it.)*

Common program metrics: (*Identify the primary program metric that corresponds to this program. If applicable, select a secondary program metric. If the metric is not from Appendix B of the ACC report, describe and justify it.)*

III. HOW MEASURES ARE USED IN PROGRAM OPERATIONS.

Are the preceding metrics currently in place, and are all projects expected to assess progress using these metrics? Please explain.

How frequently are or will outcome data be collected at the Federal level, and how are or will the data be used to monitor trends, spot problems, and identify promising practices?

What steps are in place or will the program take to ensure the outcome data collected and reported for program participants is high quality?

IV. EVALUATION.

Has the program been rigorously evaluated in the recent past? Are there plans for rigorous new evaluations? ***(Describe the methodology and scope of the evaluation, its duration, and its annual and total costs, if known.)***
If no to both, above, describe any impediments that prevent the agency from implementing an evaluation.

What are the key research questions previously evaluated or expected to be addressed? ***(If the evaluation is looking at different metrics than the chosen metrics above, please explain why.)***

Under which tier of the hierarchy of study designs do recent, ongoing, or planned evaluations fall? ***(experimental; quasi-experimental; other)***

If the program or activities do not lend themselves to study using experimental or quasi-experimental designs, describe the pathway the program will establish to ensure that the most promising practices are identified and further developed so that their impact can be rigorously evaluated in the future.

Describe how the agency will ensure that the program evaluators possess competence in evaluation methodology, subject matter expertise, and independence from the program/organization being evaluated.

Indicate whether the evaluation approach is the same or similar to one of the models in "Options to Advance Rigorous Evaluation"[4]

Indicate whether the evaluation findings were or will be applicable to the entire program or only a portion? If the latter, what proportion of the funding and what is the rationale for focusing on that portion of the program?

Describe how the agency's plans for evaluating the program have permitted or will allow for generalization of findings from the particular participants in the study to the entire program or to that portion of the program the evaluation was designed to address.

V. DISSEMINATING AND UTILIZING EVALUATION RESULTS.

Describe the agency's approach to disseminating evaluation results and highlighting effective or ineffective practices.

How does the agency use, or expect to use, evaluation findings in the design and/or operation of the program?

Describe any design or operational changes planned or recently implemented to enhance program assessment and measurement of results. ***(This can include, but is not limited to, improving data collection systems or practices, refocusing the program's mission around measurable objectives, implementing common metrics so that project performance can be compared, targeting funding to the most effective activities, disseminating information about promising practices, changing the duration of projects to enable the use of rigorous study designs.)***

---

4 Options to Advance Rigorous Evaluation

Competitive Priority. *The program gives priority consideration to award applicants that propose to conduct a scientifically-rigorous evaluation of their project. Such applicants are given additional points in the proposal evaluation process, and may also be awarded additional funds to conduct the evaluation.*

Required of all applicants. *The program requires award applicants to conduct a scientifically-rigorous evaluation of their project, and awards them additional funds to conduct the evaluation. Agency issues standards to govern quality of evaluations.*

Cross-project evaluation sponsored by the program. *The program or agency itself sponsors a scientifically-rigorous evaluation of one or more distinct interventions (e.g., a specific course curriculum) that a number of program awardees have adopted. The program or agency selects an independent researcher team to conduct this cross-project evaluation. The program requires its awardees to participate in such evaluations if asked.*

Sheltered competition. *The program sets aside a portion of its funds to conduct a "sheltered competition" for funding awards to implement a specific intervention that the program seeks to evaluate (e.g., a well-defined teacher training model that a federal teacher professional development program seeks to evaluate). The program then selects an independent research team to conduct a scientifically-rigorous evaluation of the intervention, and requires the selected awardees to participate in the evaluation.*

Waivers to allow impact study. *The agency or program waives provisions of law or regulation to allow program awardees to carry out demonstration projects of new interventions (e.g., new methods of program delivery), and in return requires such awardees to conduct a scientifically-rigorous evaluation of their demonstration project. (This option is more applicable to formula grant rather than discretionary grant programs.)*

Other

None of the above

VI. AGENCY PROGRESS SUMMARY OF ACTIONS TO IMPLEMENT ACC RECOMMENDATIONS.

Please highlight the most significant activities your agency has undertaken or has planned for the next fiscal year which address ACC recommendations. (Examples of areas in which you may have activities to summarize are given below.) Because this summary is intended to highlight the agency's overall efforts, it should not duplicate the detailed discussion in the evaluation template. ***(Agencies should limit their summaries to 2 to 3 pages).***

Note: Please be concise and factual in your responses. The final report to the President will draw upon the information you provide, and the completed templates for individual agencies may be published as appendices to the report.

Common metrics: How is the agency implementing the ACC metrics to ensure that all partners work toward common outcome goals?

Evidence-based focus: What steps is the agency taking to strengthen evaluation rigor, improve dissemination of proven practices, or modify program designs or operations to enable decision-making at the program or project level to be guided by evidence of impact?

Coordination: What is the agency doing to improve coordination with other Federal agencies with decision-makers at the State, local, or school level in ways that are likely to enhance program impact?

Appendix C
Improving the Rigor of Agency STEM Education Evaluations

The importance of STEM education: One of the goals of the Education Subcommittee is to encourage improved evaluation of STEM education programs in order to advance evidence-based policy and practices. The Subcommittee shares the concerns regarding the STEM workforce in the federal government and, more generally, the STEM skills of the U.S. workforce that have commanded much public attention recently. Because of these concerns, a number of federal agencies have invested in STEM education programs that fund projects that range from pre-K, through to post-graduate research training, and into informal education for the public. Each agency wants to wisely invest its limited resources effectively and has made a commitment to conducting rigorous evaluations.

Successful, large-scale efforts to improve STEM education are unlikely to arise *ex nihilo*. Ideally, these efforts should be based on knowledge generated from many educational research studies that follow a line of inquiry over a period of years. Government support can be crucial in the early stages of development where fundamental knowledge generation occurs. Some federal agencies fund basic research to improve our understanding of student learning in general, and mathematics and science learning in particular. Some agencies support research and development designed to apply what we already know about teaching and learning in order to create better educational interventions. Many of these interventions have specific educational goals that relate to the unique scientific mission of the agencies (e.g., laboratory internships, research training, workforce development, etc.)

Evaluating STEM education programs: It is important to choose evaluation methods that are appropriate to the research questions being asked and to each agency program's stage of development. The 2002 National Research Council (NRC) report *Scientific Research in Education* suggests, "Methods can only be judged in terms of their appropriateness and effectiveness in addressing a particular research question."[i] Methods also need to be appropriate to the stage of development in the particular research project and to the stage of theoretical development in the larger genre of work in which the particular study is conducted.

Federal STEM programs involve projects at many different stages of development; for some innovation and initial prototype development are the goal; in others, scalability and impact need to be evaluated. The Education Subcommittee certainly encourages agencies to examine "what works," but is also keenly interested in ensuring that questions of "why it works" and "what appears not to be working and why," are pursued, in order to build the educational knowledge base.

Scientifically valid education evaluations can employ a range of methodologies. These include, but are not limited to, methods for: producing descriptive summaries of a project's operation including participant viewpoints; isolating possible relationships among variables in project implementation and outcomes; and estimating the impact of particular projects.

Scientifically valid education evaluation can address a range of questions. These include but are not limited to questions about: the degree to which a program is implemented as intended; the characteristics and attitudes of the potential and actual customers of the program; the progress of program participants over time; the extent to which the program is being utilized; and the impact of the program on particular outcomes.

It is worth noting that, with appropriate modifications, all of the questions and methodologies mentioned above can be directed at either program-level (i.e., agency-level) or project-level (i.e., grantees) evaluations. Federal agencies will want to conduct both rigorous program-level evaluations and more rigorous project-level evaluations by their grantees.

Defining an intervention: Within the context of evaluating Federal STEM programs, the Education Subcommittee defines an "intervention" as the factor or factors that are under the control of and provided by a Federal agency or its agents in anticipation of affecting STEM-related outcomes. Interventions are sometimes called "treatments" or "independent variables". Interventions can be as broad as a funding stream, e.g., Federal funding to public schools in proportion to their number of low-income children with the goal of enhancing academic outcomes for children in those schools. Interventions can be as focused as encouraging a group of teachers to view a particular video of a master teacher responding to students' errors in solving mathematics problems with the hope that teachers will modify their pedagogy. It is important to distinguish the factors that are under the control of and provided by the federal agency or intervener, providing money in the first and access to the video in the second example, from the factors and experiences, often variable in nature that flow from the intervention and may mediate the outcomes that are anticipated.

Suppose a Federal agency runs a discretionary grant program in which postsecondary institutions can apply for project funds to enhance their effectiveness in retaining students in their undergraduate engineering courses of study. Individual grantees under this program may differ substantially in the type and mix of activities they support with their grant funds. Some may emphasize tutoring, others mentoring, others changes in classroom pedagogy, other changes in curriculum, and so forth. The particular mix of activities in each project could well be critical to the degree of institutional success in retaining engineering students, and should be well documented in a strong evaluation study. But for the purpose of evaluating the Federal program, the intervention is receipt of funds under the discretionary grant program not the particular way those funds are spent on a given campus. Of course a given campus or a federally sponsored research team could evaluate a particular retention project in relation to one or more alternatives, in which case the particular project would be the intervention and not the receipt of funds. The question of what is being evaluated, i.e., defining the intervention, is critical to the design of an evaluation, the selection of measures, and ultimately, rigor.

Funding streams or the receipt of grants can be evaluated as interventions and may be the most appropriate definition of the intervention for some federal programs. Other programs that are better defined in terms of particulars can also be evaluated as interventions, and evaluations of such well-defined programs are often easier to carry out because they involve an implicit or explicit theory of action that generates hypotheses about the moderators of the program's effectiveness. For example, an adjunct teacher program that makes competitive grants to to create opportunities for professionals with STEM disciplinary expertise to teach secondary school courses is based on the hypothesis that the superior content knowledge of these professionals is directed related to their effectiveness as teachers. An evaluation of this program might examine whether variations in the disciplinary content knowledge and pedagogical content knowledge of the adjunct teachers are related to differences in student outcomes. How a particular intervention is defined, guides the selection of what to measure in addition to the major outcomes of interest.

But in all cases, being clear about the question being asked is critical. Thus an evaluation that demonstrates that a particular postsecondary retention project is effective is largely irrelevant to the question of whether the Federal grant *program* (e.g., one in which all awards are aimed at

improving postsecondary retention) is effective. Likewise the evaluation that demonstrates that the recipients of grants under the Federal funding *program*, on average, increase their retention rates is largely irrelevant to the question of whether a particular retention *project* created and used on a single campus is effective. Both questions are important and potentially worth evaluating, but they would lead to very different evaluation designs.

The use of randomized controlled trials in evaluation: The randomized controlled trial (RCT), when appropriate and correctly implemented, is the most powerful design for detecting the treatment effect (impact) of an intervention. Other designs can provide useful evidence of treatment effect, but a technically sound RCT provides more assurance that the assignment of participants to treatment conditions is independent of the pretreatment characteristics of group members; thus differences between groups can be attributed to treatment effects rather than to the pretreatment characteristics.

Other methods within a family of quasi-experimental designs can be utilized to estimate an intervention's impacts when RCTs are not feasible. Unlike RCTs, they typically do not eliminate all plausible competing explanations for the obtained results. For example, in a quasi-experimental design every effort will be made carefully match the treatment and control groups on those criteria that the investigator believes are likely to affect the outcomes (e.g., socio-economic status, grade point average, ethnic group, etc.) However, quasi-experimental designs are open to the risk that the experimental and treatment groups differed in some other significant way, unknown to the investigator, which affects the experimental outcome.

Sophisticated statistical regression techniques can also be used to strengthen or weaken hypotheses about the effects of program participation. Such techniques typically produce substantially more uncertainty about the causal effects of an intervention than well-designed quasi-experiments, which, in turn, are less certain than well-designed RCTs. Thus for impact questions there is a hierarchical dimension of rigor defined by the degree to which the method eliminates explanations for the results that compete with the hypothesis that participation in the intervention is responsible. Assuming that each of these experimental designs is executed with the same degree of technical precision, RCTs sit at the top of this hierarchy.

However, some research programs will never be candidates for RCTs. Basic research and development activities do not lend themselves to RCTs nor are they at a state of maturity where RCTs would be warranted. While RCTs can be used to test the effectiveness of the mature interventions, RCTs are much less helpful in the development of the intervention itself. Other education programs cannot feasibly be evaluated over the short term with any method that compares participants with non-participants. These include programs that intend to increase the supply of something (e.g., scientific breakthroughs) or to generate a particular product (e.g., assessments of mathematics in grades 3-8 in every state). Rigorous evaluations of such programs depend on articulation of clear program goals and measure progress towards them. Similarly, because of their highly specialized, individualized nature and interrelationship with the day-to-day conduct of research, graduate and postdoctoral education programs are not well suited for randomization. At the same time, however, evaluators are increasingly turning to regression discontinuity and other appropriate quasi-experimental designs to assess graduate and postdoctoral programs.

These challenges and exceptions notwithstanding, it is the goal of the Education Subcommittee to foster and encourage federal STEM education programs to translate the best available basic research into STEM learning, to carefully design programs and intervention strategies around target audiences and program goals, and to eventually evaluate these interventions with the

most rigorous evaluation methods that are practical and appropriate for assessing program impact.

Rigor in science: In science, the term "rigor" is typically used to indicate a judgment regarding the quality of a particular investigation. Rigor is not an inherent quality that is inextricably attached to a specific research methodology. It is correct to say that a carefully implemented randomized controlled trial (RCT) will allow researchers to rigorously draw statistically based causal inferences. However, RCTs can be conducted inappropriately or with serious flaws in execution or analysis such that the term "rigor" would not apply. Similarly, clinicians can meticulously document important new findings using case studies in a way that would warrant the use of the term rigorous. A famous example is John Snow's case study in 1831 that explained the source of a cholera epidemic.[ii] This work led to public health changes that have saved millions of lives and form the basis for modern epidemiological studies, including those directed at a potential pandemic of bird flu. In each case, the term rigor refers more to the disciplined application of reason and the *appropriate* use of research methodologies to the investigation at hand than it does to any particular research methodology employed.

The Education Subcommittee defines an evaluation to be rigorous if it exhibits the following characteristics:

- The methodology aligns with the goals of the project or program being evaluated and the questions the evaluation proposes to answer.
- The evaluation strictly adheres to professionally accepted protocols of design, data collection, and data analysis.
- The data collection instruments are appropriate, reliable, and valid.
- The statistical analyses are appropriate and done correctly.
- The conclusions drawn are supported by the data and its analysis.

Clearly then, rigor must be interpreted in context. For example, attitudinal surveys and quasi-experiments vary substantially in technical quality. In the former case, issues such as whether the sample is representative and the appropriate construction of questions loom large, while in the latter case, the degree to which equivalence of groups can be demonstrated at pretest is very important. Thus the features that determine the rigor of a quasi-experiment are not directly relevant to the rigor of a survey of attitudes of people who visit science museums. Conversely, the features that determine the rigor of an attitudinal survey are not relevant to the question of whether a professional development program for mathematics teachers has an impact on their teaching skills. Thus, in isolation from the question being addressed, no particular methodology can be said to be more rigorous than another.

Education research to guide policy making: Policy makers use research for sound evidence on which to guide investments, make recommendations, and base decisions. Because problems related to STEM education and the STEM workforce have become issues of national importance, many groups within our society are turning to education research seeking solutions. Some of the evidence that policy makers desire will be best attained through the appropriate use of well-designed RCT's and quasi-experiments. Policy makers may also want to know why particular interventions do and do not work, for whom, and under what circumstances, and that may require methodologies in addition to the RCT. However, to date, RCTs and methodologically strong quasi-experiments have been used much less frequently in education research than in medicine. The reasons for this are both historical and practical.

Historically, education research is at a much earlier stage of development than is medicine. Many areas in education lack the firm theoretical underpinnings that guide modern medical research. (Some of the federal STEM education effort supports research that is helping fill those gaps in theory.)

On a practical level, designing controlled trials in educational settings presents unique challenges to the researcher.[iii] It is probably not possible to design a "double-blind" trial in a real-world education setting (where students, teachers, and researchers are all unaware of the intervention being tested). This is a serious theoretical impediment to the "internal validity" of the trial because it becomes difficult to control for the changes in behavior that might result in researchers or subjects once they know the treatment group to which they have been assigned. However, the strength of the self-fulfilling prophecy or expectation effect that might be introduced by students or teachers knowing that they are in an "intervention" is known to be weak for many academic outcomes. And when the outcomes of interest are potentially subject to a self-fulfilling prophecy, e.g., attitudes, the standard methodological control is to compare two interventions, both of which are presented in equally positive terms, and to directly measure and control statistically for any group differences in expectations for success. This problem has been well worked out in the behavioral sciences, e.g. Research Design in Clinical Psychology, Alan Kazdin, (2002).

The quality of an RCT can be judged using several criteria, some of more interest to policy makers than others. These criteria include *internal validity*, *external validity*, and *construct validity*.

Internal Validity: Internal validity is the confidence one has that the outcome observed in the trial was actually the result of the intervention(s) being tested. The greater the methodological rigor of the study, the higher will be the confidence in the conclusions of that study. Note that achieving increasing levels of confidence will still never allow for the attainment of "certainty." However, in well-designed studies, where the internal validity is high, it is possible to use statistical methods to calculate the likelihood that the observed outcomes are due to the intervention or, with some stated probability, to chance alone. Some of the factors that can influence internal validity include *sample size* (whether the subjects are sufficiently numerous to detect the effect of the treatment), *compliance* (whether the different treatment groups faithfully implement their assigned treatments), *effectiveness of randomization*, *attrition* (loss of subjects from control or treatment groups), and *contamination* (whether uncontrolled factors influence only the control or treatment groups).

External Validity: External validity relates to the generalizability of the outcome to other groups, other locations, and other times. External validity is typically the characteristic of greatest interest to policy makers who hope to know whether a particular intervention will work in their own community. A number of factors can affect external validity, including (1) whether the intervention itself can be reproduced accurately in another time or place and (2) the likelihood that the intervention will result in the same outcomes when transported to a new time or place.

External validity is a major challenge in educational research. It is more difficult, time consuming, and expensive to rigorously demonstrate external validity than internal validity. External validity can be generated and enhanced in two ways. The first is replication, i.e., separate studies conducted at different times and places that produce similar findings. The second is large-scale single studies that sample a wide variety of settings and circumstances. External validity is always dimensional, and more is always better.

Construct Validity: Construct validity related to measurement methods refers to the ability to know that you are assessing a particular attribute accurately. For example, how do we measure student achievement or understanding in science? Are the state, national, or international assessment exams accurate measures of them? In the current national conversation concerning "competitiveness," business organizations are asking for students who are better *problem solvers*, *critical thinkers*, and *innovators*. How confident are we that we have, and are using, good measures to assess these skills?

Conclusions on rigor and program evaluation: Research and evaluation methodologies should be appropriately matched to the scientific question being asked, and multiple methods are needed to build and advance the knowledge base. While we want to encourage increased use of RCTs in education research and evaluation, establishing causality in science is a complex undertaking (and education science is no exception). Establishing causality requires a coherent body of theory that can be used to predict specific relationships between program interventions and student outcomes. Even then, a single small RCT, no matter how well designed, will not be sufficient to establish general conclusions that a program or intervention works. The ability to generalize findings requires a large extent of evidence gathered in different settings and circumstances. Expert review of the accumulated research, carefully considering the validity and relevant characteristics of each study to substantiate the results and outcome, will also be a necessary step.

---

[i] National Research Council. *Scientific Research in Education*. Washington, D.C.: National Academies Press, 2002.

[ii] Hemple, S. *The Strange Case of the Broad Street Pump: John Snow and the Mystery of Cholera.* Berkeley, CA: University of California Press, 2007.

[iii] Brass, C.T., Nunez-Neto, B., and Williams, E.D. *Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues.* CRS Report RL33301, Congressional Research Service, Library of Congress, 2006.

Appendix D
Key elements of the evaluation plans for ACC Agencies' High Leverage Programs