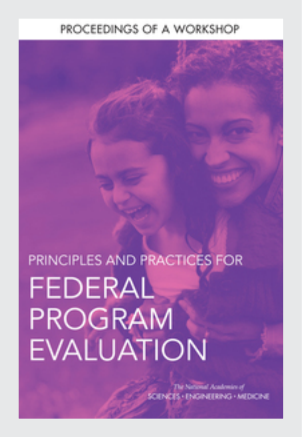

Principles and Practices for Federal Program Evaluation: Proceedings of a Workshop

DETAILS

66 pages | 6 x 9 | PAPERBACK
ISBN 978-0-309-46275-4 | DOI 10.17226/24831

CONTRIBUTORS

Jordyn M. White, Rapporteur; Steering Committee on Principles and Practices for Federal Program Evaluation; Committee on National Statistics; Division of Behavioral and Social Sciences and Education; National Academies of Sciences, Engineering, and Medicine

PRINCIPLES AND PRACTICES FOR FEDERAL PROGRAM EVALUATION

PROCEEDINGS OF A WORKSHOP

Jordyn M. White, *Rapporteur*

Steering Committee on Principles and Practices
for Federal Program Evaluation

Committee on National Statistics

Division of Behavioral and Social Sciences and Education

The National Academies of
SCIENCES · ENGINEERING · MEDICINE

THE NATIONAL ACADEMIES PRESS
Washington, DC
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, NW Washington, DC 20001

This activity was supported by contract number HHSP233201400020B between the U.S. Department of Health and Human Services: Administration for Children and Families (ACF) and the Office of the Assistant Secretary for Planning and Evaluation (ASPE); U.S. Department of Labor (DOL); Institute for Education Sciences (IES); U.S. Office of Management and Budget (OMB); and the National Academies of Sciences, Engineering, and Medicine. Support for the work of the Committee on National Statistics is provided by a consortium of federal agencies through a grant from the National Science Foundation (award number SES-1024012). Any opinions, findings, conclusions, or recommendations expressed in this publication do not necessarily reflect the views of any organization or agency that provided support for the project.

International Standard Book Number-13: 978-0-309-46275-4
International Standard Book Number-10: 0-309-46275-4
Digital Object Identifier: https://doi.org/10.17226/24831

Additional copies of this publication are available for sale from the National Academies Press, 500 Fifth Street, NW, Keck 360, Washington, DC 20001; (800) 624-6242 or (202) 334-3313; http://www.nap.edu.

Printed in the United States of America

Suggested citation: National Academies of Sciences, Engineering, and Medicine. (2017). *Principles and Practices for Federal Program Evaluation: Proceedings of a Workshop*. Washington, DC: The National Academies Press. doi: https://doi.org/10.17226/24831.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

The **National Academy of Sciences** was established in 1863 by an Act of Congress, signed by President Lincoln, as a private, nongovernmental institution to advise the nation on issues related to science and technology. Members are elected by their peers for outstanding contributions to research. Dr. Marcia McNutt is president.

The **National Academy of Engineering** was established in 1964 under the charter of the National Academy of Sciences to bring the practices of engineering to advising the nation. Members are elected by their peers for extraordinary contributions to engineering. Dr. C. D. Mote, Jr., is president.

The **National Academy of Medicine** (formerly the Institute of Medicine) was established in 1970 under the charter of the National Academy of Sciences to advise the nation on medical and health issues. Members are elected by their peers for distinguished contributions to medicine and health. Dr. Victor J. Dzau is president.

The three Academies work together as the **National Academies of Sciences, Engineering, and Medicine** to provide independent, objective analysis and advice to the nation and conduct other activities to solve complex problems and inform public policy decisions. The National Academies also encourage education and research, recognize outstanding contributions to knowledge, and increase public understanding in matters of science, engineering, and medicine.

Learn more about the National Academies of Sciences, Engineering, and Medicine at **www.nationalacademies.org**.

The National Academies of
SCIENCES • ENGINEERING • MEDICINE

Consensus Study Reports published by the National Academies of Sciences, Engineering, and Medicine document the evidence-based consensus on the study's statement of task by an authoring committee of experts. Reports typically include findings, conclusions, and recommendations based on information gathered by the committee and the committee's deliberations. Each report has been subjected to a rigorous and independent peer-review process and it represents the position of the National Academies on the statement of task.

Proceedings published by the National Academies of Sciences, Engineering, and Medicine chronicle the presentations and discussions at a workshop, symposium, or other event convened by the National Academies. The statements and opinions contained in proceedings are those of the participants and are not endorsed by other participants, the planning committee, or the National Academies.

For information about other products and activities of the National Academies, please visit www.nationalacademies.org/about/whatwedo.

STEERING COMMITTEE ON PRINCIPLES AND PRACTICES FOR FEDERAL PROGRAM EVALUATION

GROVER J. WHITEHURST (*Chair*), The Brookings Institution, Washington, DC
JUDITH M. GUERON, MDRC, New York
REBECCA A. MAYNARD, Graduate School of Education, University of Pennsylvania
MARTHA MOOREHOUSE, Independent Consultant, Los Altos, CA
HOWARD ROLSTON, Independent Consultant, Arlington, VA
WILLIAM J. SABOL, Westat, Inc., Rockville, MD

JORDYN M. WHITE, *Study Director*
CONSTANCE F. CITRO, *Board Director*
CYNTHIA THOMAS, *Senior Program Officer*
MICHAEL SIRI, *Program Coordinator*

COMMITTEE ON NATIONAL STATISTICS

ROBERT M. GROVES (*Chair*), Department of Mathematics and Statistics and Department of Sociology, Georgetown University
FRANCINE BLAU, Department of Economics, Cornell University
MARY ELLEN BOCK, Department of Statistics, Purdue University
ANNE C. CASE, Woodrow Wilson School of Public and International Affairs, Princeton University
MICHAEL E. CHERNEW, Department of Health Care Policy, Harvard Medical School
JANET CURRIE, Woodrow Wilson School of Public and International Affairs, Princeton University
DONALD A. DILLMAN, Department of Sociology, Washington State University
CONSTANTINE GATSONIS, Center for Statistical Sciences, Brown University
JAMES S. HOUSE, Survey Research Center, Institute for Social research, University of Michigan
THOMAS L. MESENBOURG, U.S. Census Bureau (Retired)
SUSAN A. MURPHY, Department of Statistics, University of Michigan
SARAH M. NUSSER, Department of Statistics, Center for Survey Statistics and Methodology, Iowa State University
COLM A. O'MUIRCHEARTAIGH, Harris Graduate School of Public Policy Studies, The University of Chicago
JEROME P. REITER, Department of Statistical Science, Duke University
ROBERTO RIGOBON, Sloan School of Management, Massachusetts Institute of Technology
JUDITH A. SELTZER, Department of Sociology, University of California, Los Angeles
EDWARD H. SHORTLIFFE, Columbia University and Arizona State University

BRIAN HARRIS-KOJETIN, *Director*

Acknowledgments

This Proceedings of a Workshop was reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise. The purpose of this independent review is to provide candid and critical comments that will assist the National Academies of Sciences, Engineering, and Medicine in making each published proceedings as sound as possible and to ensure that it meets the institutional standards for quality, objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential to protect the integrity of the process.

We thank the following individuals for their review of this proceedings: Ruth Levine, Global Development and Population Program, The William and Flora Hewlett Foundation, and Mark D. Shroder, Research, Evaluation, and Monitoring, U.S. Office of Housing and Urban Development.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the proceedings nor did they see the final draft before its release. The review of this proceedings was overseen by Janet Currie, Department of Economics and Center for Health and Well-Being, Woodrow Wilson School of Public and International Affairs, Princeton University. She was responsible for making certain that an independent examination of this proceedings was carried out in accordance with standards of the National Academies and that all review comments were carefully considered. Responsibility for the final content rests entirely with the rapporteur and the National Academies.

Contents

APPENDIXES

1

Introduction

This proceedings summarizes the presentations and discussions at the 1-day public workshop on Principles and Practices for Federal Program Evaluation, which was held in Washington, DC, in October 2016. The workshop was organized as part of an effort to assist several agencies: in the U.S. Department of Health and Human Services, the Office of the Assistant Secretary for Planning and Evaluation (ASPE) and Administration for Children and Families (ACF); in the U.S. Department of Labor, the Office of the Chief Evaluation Officer (CEO); and in the U.S. Department of Education, the Institute of Education Sciences (IES). The purpose of the workshop was to consider ways to bolster the integrity and protect the objectivity of the evaluation function in federal agencies—a process that is essential for evidence-based policy making. The scope of the workshop included evaluations of interventions, programs, and practices intended to affect human behavior, carried out by the federal government or its contractual agents, that result in public reports sponsored by the federal government and are intended to provide information on their impacts, cost, and implementation.

BRINGING FEDERAL EVALUATION TO THE FOREFRONT

The federal government has taken several steps over the past two decades to bolster the credibility of scientific evidence. The Information

Quality Act of 2001[1] and the Office of Management and Budget's (OMB) *Guidelines for Ensuring and Maximizing the Quality, Objectivity, Utility, and Integrity of Information Disseminated by Federal Agencies*[2] both advise agencies on preserving the quality of data from collection through dissemination and on developing the appropriate administrative mechanisms to carry out these standards. OMB's *Guidelines* also provide guidance on remaining true to the intended users and uses of the data (utility) while presenting the data in a clear, transparent, and unbiased manner (objectivity) that is free from corruption or undue influence (integrity).

In its chapter entitled "Building the Capacity to Produce and Use Evidence," the Analytical Perspectives Component of the Budget of the United States Government, Fiscal Year 2017 emphasized the importance of establishing centralized or chief evaluation offices in federal agencies and supported the development of guidelines for federal program evaluations, stating that "Many Federal evaluators believe that establishing a common set of government-wide principles and practices for evaluation offices could help to ensure that Federal program evaluations meet scientific standards, are designed to be useful, and are conducted and the results are disseminated without bias or undue influence." The document went on to highlight five fundamental principles in developing standards for evaluation: rigor, relevance, transparency, independence, and ethics.

ACF and CEO have both issued evaluation policy statements for their organizations that address those principles,[3,4] and the Department of Education issued a 9-page departmental directive in 2014 that addressed scientific integrity in research activities departmentwide.[5] These three policies were among the documents provided to participants for review in advance of the workshop.

DRAWING UPON EXAMPLES FROM THE FEDERAL STATISTICAL SYSTEM

In the past several years, the heads of many federal agencies have articulated a long-range goal to build an infrastructure to strengthen and guide federal evaluations—one that would promote continuity and support for certain high-level principles and practices across agencies and chang-

[1] See Section 515 of the Treasury and General Government Appropriations Act, 2001 (Pub. L. No. 106-554, 44 U.S.C. § 3516 note).

[2] See https://obamawhitehouse.archives.gov/omb/fedreg_final_information_quality_guidelines [May 2017].

[3] See https://www.acf.hhs.gov/sites/default/files/opre/acf_evaluation_policy_november_2013.pdf [May 2017].

[4] See https://www.dol.gov/asp/evaluation/EvaluationPolicy.htm [May 2017].

[5] See https://www.state.gov/s/d/rm/rls/evaluation/2015/236970.htm [May 2017].

ing federal administrations. Attention to evaluations and the level of the agency responsible for such activities vary considerably across the federal government.

As with federal evaluation, the U.S. federal statistical system itself is highly decentralized; while statistical activities are conducted in more than 100 agencies, only a few focus on producing statistics as part of their primary mission. In her introductory remarks, Constance Citro (Committee on National Statistics [CNSTAT]) discussed the history of CNSTAT and its joint efforts with OMB to facilitate coordination and collaboration across the statistical system. To provide advice to Congress and the Executive Branch on establishing a new statistical agency and describe foundational principles for its activities, CNSTAT published *Principles and Practices for a Federal Statistical Agency* (National Research Council, 1992), and since 2001 the volume has been updated every 4 years (at the beginning of a new administration or second term). Citro said it is widely recognized to have been helpful in preserving the independence of federal statistical agencies. It is intended to bolster statistical practices from undue partisan or political influence, and she noted how that goal aligned with the sponsors' goal for this workshop.

IMPETUS FOR A WORKSHOP AND WORKSHOP CHARGE

To further the long-range goal of strengthening federal evaluations, heads of several federal evaluation offices arranged for CNSTAT to convene a 1-day planning meeting, which was held in September 2015, to assess the usefulness of developing a document for federal evaluation programs modeled after *Principles and Practices for a Federal Statistical Agency*: a high-level set of guidelines that would help evaluation offices maintain standards for their programs across administrations and changes in political-level personnel. At the planning meeting, the cognizant federal agencies decided that a public workshop, with full discussion of existing policies for federal program evaluations and consideration of issues in building on these policies, would be a useful next step. The workshop was to be designed so as not to prejudge the value of a volume along the lines of *Principles and Practices for a Federal Statistical Agency*.

The charge to CNSTAT was to organize a workshop to consider ways to strengthen existing federal evaluation policies and to institutionalize the principles: see Box 1-1 for the full statement of task. To address the charge, CNSTAT worked with ACF, ASPE, CEO, and IES to form the Steering Committee on Principles and Practices for Federal Program Evaluation. The goal of the workshop was to review and comment on existing federal policies, which generally reference such principles as rigor, relevance, transparency, independence, and ethics, as well as objectivity, clarity, reproduc-

BOX 1-1
Statement of Task

An ad hoc steering committee will be appointed to organize a one-day, public workshop to comment on existing high-level guidelines for federal program evaluation and consider issues in developing a broader document, which might follow the model of the CNSTAT publication, *Principles and Practices for a Federal Statistical Agency.* The steering committee will invite participants to review and comment on existing agency policies for federal program evaluation, which generally reference such principles as rigor, relevance, transparency, independence, and ethics, as well as objectivity, clarity, reproducibility, and usefulness. The steering committee will develop an agenda for the workshop, including sessions that assess existing documents, consider ways to build on existing documents, including drawing lessons from CNSTAT's experience in developing *Principles and Practices for a Federal Statistical Agency*, and identify issues for agencies to consider should they decide to develop a P&P for federal program evaluation or take other steps to bolster the integrity and protect the objectivity of the evaluation function. A rapporteur will prepare a proceedings of the workshop consistent with Academies' institutional guidelines.

ibility, and usefulness, and to discuss the potential for developing a broader policy document.

ORGANIZATION OF THE VOLUME

This proceedings describes the workshop presentations and discussions that followed each topic: see workshop agenda in Appendix A. Chapter 2 presents the history of federal program evaluation and its successes and challenges from a variety of perspectives. Chapter 3 explores several prominent evaluation shops and their approaches to protecting the integrity and objectivity of evaluation work, with presentations from ACF, CEO, IES, and the Millennium Challenge Corporation. Chapter 4 covers the session in which workshop participants were invited to share their insights on the policies and policy-making processes at their respective agencies and organizations. The discussion in Chapter 5 focuses on the components necessary to advance high-quality evaluations and protect the infrastructure that supports them. In Chapter 6 a former statistical agency head and a former key player at OMB share their experiences institutionalizing the federal statistical system and its implication for developing a similar structure for evaluation. Chapter 7 addresses the need to develop guidance to support objective evaluation while simultaneously mitigating any potential resistance to the development of this type of policy. Finally, Chapter 8 discusses key themes and possible next steps in bolstering the principles and practices for federal program evaluation.

This proceedings has been prepared by the workshop rapporteur as a factual summary of what occurred at the workshop. The steering committee's role was limited to planning and convening the workshop. The views contained in the report are those of individual workshop participants and do not necessarily represent the views of all workshop participants, the steering committee, or the National Academies of Sciences, Engineering, and Medicine.

2

The Evolution of Federal Evaluation

Christine Fortunato (Administration for Children and Families [ACF]) followed Connie Citro's introductory remarks by emphasizing the importance of the workshop, stating that developing an infrastructure to guide federal evaluations and support high-level principles across offices and administrations has been a long-term goal of several federal agencies. She said that such an infrastructure would also help ensure that the programs are conducted and the results disseminated without bias or undue influence. Fortunato said that the federal government has taken several steps to foster the credibility of scientific evidence, including Statistical Policy Directive 1 from the U.S. Office of Management and Budget (OMB),[1] the Information Quality Act,[2] and the creation of specific evaluation policy statements by several federal agencies. Several of the agencies' documents prescribe the core principles of rigor, relevance, transparency, independence, and ethics, which she explained would be focal points of the workshop discussion.

FEDERAL EVALUATION, WITH THICK SKIN

Steering committee chair Grover "Russ" Whitehurst stressed how essential it is for the federal government to have a strong evaluation effort, marked by rigor and independence, enabling agencies to provide accurate

[1] *Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units.* Available: https://www.federalregister.gov/documents/2014/12/02/2014-28326/statistical-policy-directive-no-1-fundamental-responsibilities-of-federal-statistical-agencies-and [May 2017].

[2] Section 515 of the Consolidated Appropriations Act, 2001 (Pub. L. 106–554).

and timely information to decision makers. Using an anecdote about an early federal evaluation in which he was involved, Whitehurst described how the results of federal evaluation are often received with trepidation and even embarrassment when they question the effectiveness of a program or refute the desired outcome or popular choice. In 2005, a randomized controlled trial conducted on the 21st Century Community Learning Centers After School Program found that the program did not improve participants' academic skills and actually increased their misbehavior, e.g., being suspended from school. This news came much to the chagrin of then-Senator Arlen Specter, then-governor Arnold Schwarzenegger, and a host of other program grantors, community members, and advocates. They questioned the quality and relevance of the evaluation rather than accepting the importance of its findings to making decisions about the future direction of the program.

Whitehurst said that while federal evaluation is a more mature, more secure, and less isolative endeavor than it was in previous years, the field still has a long way to go. He cited a 2013 report from the U.S. Government Accountability Office (GAO): it found that less than half of 24 agencies surveyed conducted any evaluations at all of their programs, and only 7 had a centralized leader with responsibility for overseeing evaluation activities. Another issue, Whitehurst noted, is that the evaluations often do not reach their most important audiences. According to that same GAO report, more than half of senior government leaders had no experience with evaluation of the programs for which they were responsible. He encouraged participants to consider the history of federal program evaluation, its current status, and the possible development of more formal principles and practices.

HOW FAR WE HAVE COME: "THE INEVITABLE MARCH OF SCIENCE" OR AN ONGOING STRUGGLE?

Moderator Howard Rolston (member, steering committee) began the session on the history of federal program evaluation by praising the field for the progress that has been made over the past 50 years. He noted that the continuous growth, progress, and improvements in evaluation can lead observers to think these advances are simply a product of "the inevitable march of science": moving forward, learning more, and progressing through innovation. He cautioned, however, not to take the progress for granted, because although the overall trend has been more and better evaluations, there have been setbacks and points at which the field of evaluation has come under threat. Rolston noted that much of the past decades' progress has resulted from the individual efforts of committed federal staff, philanthropic funders, committed academics, and advocates for evidence-based policy. He has seen the trend move toward creating institutionalized structures that can protect the quality and dissemination of evaluation findings—one element of that trend being the formulation of evaluation principles with practices in place to support them.

Larry Orr (Bloomberg School of Public Health, Johns Hopkins University) started his presentation by sharing the two overarching questions the panelists decided were most useful to address in terms of the history of evaluation: What have been the major challenges to the federal government in generating and using rigorous independent research? What circumstances over time have reduced or exacerbated vulnerabilities in evaluation work? He noted three challenges and discussed how they have evolved over time: resources for research and evaluation, resistance to rigorous evaluation, and convincing policy makers to use evaluation results.

In terms of resources, Orr said that in the 1970s, when he was director of the Office of Income Security Policy Research, then a part of the Office of the Assistant Secretary for Policy and Evaluation (ASPE) in the U.S. Department of Health and Human Services, his unconstrained operating budget was $25 million, which is essentially equivalent to $100 million in 2016. He researched and found that in 2016, ASPE's entire budget was $56 million. (He did note, however, that the decrease may be due in part to a transfer of several ASPE research responsibilities to another office, which had a 2012 operating budget of $107 million—still just roughly above his evaluation budget four decades earlier.) Orr surmised that there has been limited progress by way of increasing resources for program evaluation and that it is still grossly underfunded by an "order of magnitude." By comparison, he mentioned how medical researchers spend $30 billion to conduct 10,000 clinical trials (a form of evaluation) a year; in contrast, in social policy billions of dollars are spent on programming but significantly less on finding out whether or not those programs work.

Orr said that he sees resistance to rigorous evaluation both in the government and, surprisingly, in the research community. He cited *Fighting for Reliable Evidence* (Gueron and Rolston, 2013) as an account of the challenges of incorporating random assignment experimentation in social policy research. He also discussed how *Equality of Educational Opportunity* (Coleman, 1966) essentially turned the education community against quantitative evaluation for several decades. The prevailing theory at the time was that in order to understand education one had to look on a micro level at the success of individual school systems, which is not helpful in setting national policy. The establishment of the Institute of Education Sciences in 2002 changed that rationale, and its What Works Clearinghouse now has identified nearly 600 well-conducted randomized trials on education programs and practices. Orr noted that the field of international development also initially resisted evaluation, but since 2000 there have been approximately 1,700 randomized controlled trials in developing countries:

770 of those were conducted by the Poverty Action Lab at the Massachusetts Institute of Technology, whose mission is "to reduce [international] poverty by ensuring that policy is informed by scientific evidence."[3] He sees this as a clear indication of progress for evaluation.

Convincing policy makers to act on research results is one of his biggest challenges in the field, Orr said. He reminded the group that evaluation is only one of the many factors that influence policy, and it currently plays a very small part. He mentioned how in *Show Me the Evidence: Obama's Fight for Rigor and Evidence in Social Policy* (Haskins and Margolis, 2015) there is a pie chart depicting all of the factors that influence social policy. Among the categories, which include advocacy groups, committee staff, and news media, the slice for research is one of the smallest, at just 1 percent. Orr believes, however, that the role of evaluation will continue to grow because of an increasing number of congressional mandates for rigorous research and because of the establishment of rigorous analysis as a standard in policy making by the Congressional Budget Office. He also noted OMB's efforts toward the increased use of rigorous evidence—namely, the Bush administration's PART (Program Assessment Rating Tool) and President Obama's evidence-based policy initiatives.

Jean Grossman (Princeton University and MDRC) spoke about the challenges she faced both from within the federal government as an evaluation officer and in her role as a federal contractor. She noted three main issues: politics, money, and regulations. She said that politics is "the elephant in the room," and evaluators are constantly fighting political pressure and a reluctance to hear or release the results of an evaluation that do not align with the program's original expectations. She said many policy makers and others view evaluation as a way of determining whether or not a program works, when it is more about determining whether or not a program works better than something else. Grossman remembers wishing when she was the chief evaluation officer at the Department of Labor that there had been a safe way or space in which to conduct evaluations without the looming fear of defunding—where the environment was centered more on continuous improvement.

As an evaluator, Grossman recalled instances in which results were not released if they did not reinforce expectations, and she even on rare occasions felt pressure from a funder to reword an evaluation to better align with the current policy. She believes that not releasing reports happens less now that evaluation agencies are registering their evaluations and publicize their reports' due dates. Political pressure can occasionally prove to be beneficial, however, when it is used to inquire about evaluation and ask for the public release of information.

[3] See www.povertyactionlab.org [May 2017].

Political timing is another factor, Grossman said, since the average 4-year time horizon for most policy makers often requires that programs be evaluated in that time frame, which is often too short a time from their inception for a meaningful evaluation. With all the iterative changes that occur in the initial months of a program's launch, the services provided to participants in a randomized sample may change from those that were originally planned to be evaluated.

With regard to money, Grossman pointed out that only a small subset of federal funds goes directly to program evaluation—sometimes less than 0.5 percent for an agency. It is usually the case that programs do not get evaluated unless money has been earmarked or set aside specifically for that purpose. Seeking approval to use other administrative funding for evaluation can be difficult when, as she noted, the sentiments around evaluation and its uses are often negative.

Grossman described how regulations also add constraints to program evaluation—the biggest one she faced was the OMB Paperwork Reduction Act.[4] While the target turnaround for OMB approval is 3 months, the average is 7-9 months. Considering that it may take a few more months to start a program and to develop an intake form or a baseline survey, a year could elapse before participants are enrolled and staff are able to collect critical baseline information. Since most evaluation contracts last less than 5 years, this is often time that the evaluation cannot spare. As a result, the Paperwork Reduction Act makes it difficult to obtain the requisite baseline data needed for a thorough comparison and essentially limits the work that can be done. The cost and work hours required to complete an OMB clearance package can also be prohibitive for some contractors.

Ron Haskins (Brookings Institution) talked about his experience with the 21st Century Community Learning Centers: the findings from the Mathematica evaluation showed that the program did not affect student outcomes,[5] but those findings were met with much resistance, both by academics and politicians who strongly advocated for the program. Then-candidate for California governor Arnold Schwarzenegger used strong community support and anecdotal evidence to justify his stance; to Haskins, however, any sentence akin to saying "Everybody knows this program works" is an enemy of evidence-based policy.

[4] "The purpose of the Paperwork Reduction Act (PRA), which governs information collections, is to minimize paperwork, ensure public benefit, improve Government programs, [and] improve the quality and use of Federal information to strengthen decision making, accountability, and openness in Government and society": www.doleta.gov/ombcn/ombcontrolnumber.cfm [May 2017].

[5] *The $1.2 Billion Afterschool Program That Doesn't Work*. Available: www.brookings.edu/research/the-1-2-billion-afterschool-program-that-doesnt-work [May 2017].

Haskins stressed the importance of having a statute that requires an evaluation when establishing or appropriating money for a program. He said that going a step further and adding language in the statute about random assignment can also prove very useful; adding evaluation language in the Welfare Reform Act of 1996[6] improved the utility of the resulting programs and the quality of the data collected for evaluations. Similar language was also added to the Senate legislation for the World Bank. He talked about the discussions he had with Hill senior staffers and how many of them knew about random assignment when he conducted interviews for *Show Me the Evidence* (Haskins and Margolis, 2015).

To show how the conversation on evaluation has evolved in just a short time, Haskins highlighted an excerpt from the Affordable Care Act ("Obamacare," 42 U.S.C. 711) on early childhood home visiting programs, which specifically calls for evaluation through rigorous randomized controlled research designs. Haskins said that the Brookings Institution will release results from the largest evaluation to date in 2018, which will include data from home visiting programs across the country and include findings on program implementation and impacts based on the multiple randomized controlled trials. Under this same legislation, the Office of Adolescent Health released 41 evaluations of local teenage pregnancy prevention programs in fall 2016, most of which were random assignment studies.

Rolston circled back to his point about how the campaign for rigorous evaluation was initially spawned by invested individuals and their common interests, but that it has evolved into something more institutional. When he invited comments from participants, Judith Gueron (member, steering committee) emphasized the need to take the potential for future threats to program evaluation very seriously. She talked about how Reagan administration officials viewed most social science researchers as left-wing ideologues whose work was not objective or specific. As a result, they drastically reduced both program and evaluation budgets, which led to reductions in the evaluation workforce and unfinished studies. However, this reduction led to a shift to foundation-funded studies, which, because of their independence from government entities and the use of strong methodology (including randomized controlled trials), became widely accepted within both major political parties. Evaluators used an effective and unbiased communication strategy that educated various audiences and spurred community support for rigorous research. Gueron encouraged the group to keep this in mind as it considered how to fortify the principles and practices of federal program evaluation.

[6] See http://royce.house.gov/uploadedfiles/the%201996%20welfare%20reform%20law.pdf [May 2017].

Lauren Supplee (Child Trends) added two points about staffing. First, she noted that human resource challenges and regulations within the federal government can make it difficult for evaluation offices to hire quality staff. Second, she noted the need for capacity at the level of senior leadership in program offices to understand scientific evidence and be able to identify its uses and limitations. She believes that it is critical to focus on these staffing needs.

Sandy Davis (Bipartisan Policy Center) echoed Gueron's point about the effectiveness of analysis in a political setting. He added that independent evaluations in any field are not going to be successful in an intense political setting unless the evaluation offices have a known history of being objective; that objectivity can prove useful when speaking to politicians and members of Congress. Davis noted that while there often are other political forces at play when it comes to decision making, it is important that evaluations and evidence have a seat at the table. He also stressed how important it is for the evaluation to be conducted independently and in a timely manner.

Haskins agreed with Davis that there will always be people opposed to evaluation and reiterated the need to include evaluation requirements in legislation. He said that appropriations committee members and their staffs do not like to hear that a program they sponsored does not work or does not produce a major impact, yet that is the evaluation result for over 80 percent of social programs. He said the focus should be on solving a problem, not saving a program.

Rebecca Maynard (member, steering committee) spoke about her experience with evaluation of an abstinence program (Devaney et al., 2002) and how she and her colleagues used a strong technical working group to help them navigate difficult political challenges. She said that there was initial bias in the evaluation community against working on such a seemingly taboo topic, but noted that the study changed her view on research. With regards to abstinence, Maynard said it is clear that abstinence is a sure way to prevent pregnancy; the question was whether teaching abstinence without information on alternative methods, which is what the policy implied, was more effective than teaching abstinence alongside other contraception measures. Maynard said she and her colleagues designed the study so it was a "win-win"—focused on neither abstinence nor comprehensive sexual education, but instead on the differences in outcomes that resulted from the abstinence policy compared with the status quo. The focus was on the health and welfare of the children and not on the success of the program; consequently, both proponents and opponents of abstinence-only policies would have an interest in the results. She said that that kind of objectivity and respect for the design of the evaluation was imperative.

Whitehurst wrapped up the history session by summarizing what he heard as the three main points coming out of the discussion. First, it is important to include evaluation practices in legislation. Second, evaluations should be conducted with objectivity, and Congress and other stakeholders should keep their direction with regard to the purpose of an evaluation at a broad level rather than specifying detailed questions that may be impractical or impossible to answer. Third, evaluation agencies should have access to funding that is adequate to carry out high-quality evaluations that are linked programmatically so as to produce knowledge that will be useful in the long term.

3

The Standard Bearers of Federal Evaluation

ADMINISTRATION FOR CHILDREN AND FAMILIES

Naomi Goldstein (Administration for Children and Families [ACF]) provided background on ACF's evaluation policy. She noted that the agency's leaders encouraged the evaluation office to develop the policy and that the process to establish it—which included reviewing existing policies from other federal agencies, as well as the American Evaluation Association Roadmap[1]—was fairly straightforward. The policy (published in 2012) confirms the agency's commitment not only to conducting evaluations, but also to using evidence from evaluations to inform policy and practice. It was intended to clarify a few key governing principles, disseminate them both internally and externally, bolster their implementation, and protect them against potential threats.

Goldstein reminded the workshop participants that evidence is just one component of decision making, and evaluation is but one form of evidence, along with such factors as descriptive research studies, performance measures, financial and cost data, survey statistics, and program administrative data. While the ACF policy focuses primarily on evaluation, many of the principles also apply to the development and use of other types of evidence.

Goldstein discussed the five principles in ACF's policy: rigor, relevance, transparency, independence, and ethics.

Rigor means getting as close as possible to the truth and being committed to using the most appropriate methods to do so. Rigor is not restricted

[1] See http://www.eval.org/evaluationroadmap [May 2017].

to impact evaluations; it is also necessary in implementation evaluation, process evaluation, descriptive studies, outcome evaluations, formative evaluations, and in both qualitative and quantitative approaches. Goldstein noted that rigor does not automatically mean the use of randomized controlled trials—although those trials are generally considered to have the greatest internal validity, particularly for questions about cause and effect and are preferred for addressing such questions.

Rigor requires having appropriate resources, a workforce with appropriate training and experience, a competitive acquisition process, and—along with impact studies—robust implementation components that will enable evaluators to identify why a program did or didn't work or what elements were associated with greater impacts.

Relevance means setting evaluation priorities that consider many factors, including legislative requirements, the originating agency's interests, and those of other stakeholders: state and local grantees, tribes, advocates, and researchers. Relevance can be strengthened by the presence of strong internal and external partnerships and by embedding an evaluation plan into the initial program planning. It is also important to disseminate the findings in useful ways. Goldstein stressed that rigor without relevance could yield studies that are accurate but not useful.

Transparency means operating in a way that supports credibility of the findings and allows for critique and replication of the methods used in an evaluation. It promotes accessibility and reinforces a commitment to share evaluation plans in advance and release results regardless of the findings. Goldstein said evaluation reports should: describe the methods used, including strengths and weaknesses; discuss the generalizability of the findings; present comprehensive results, including unfavorable and null results; and be released in a timely manner. She noted that ACF also archives evaluation data for secondary use.

Goldstein said that independence and transparency are the protective goals of ACF's evaluation policy. They help create a culture in which broad dissemination of results becomes the standard. Independence, particularly when coupled with objectivity, is a core principle of evaluation, she said: although many parties should contribute to identifying evaluation questions and priorities, study methods and findings should be insulated from bias and undue influence.

Ethics, Goldstein emphasized, means recognizing the importance of safeguarding the dignity, rights, safety, and privacy of the participants in evaluation studies.

Goldstein closed by noting that having a policy has helped the agency clarify its goals and principles and that disseminating the policy helped make the agency's principles a shared set of values both in the organization and with its program partners.

DEPARTMENT OF LABOR

Demetra Nightingale (Urban Institute), who previously worked at the Department of Labor (DOL), began by emphasizing the importance for professional evaluators to tap into their professional networks and share their knowledge. She described DOL's mission, which includes promoting the welfare and protecting the rights of wage earners, job seekers, and retirees in the United States. She said that many of DOL's dozen or so operating agencies, such as the Employment and Training Administration and the Occupational Safety and Health Administration, have their own evaluation offices. As chief evaluation officer, her role was not to centralize evaluation, but to raise the quality of and consciousness around evaluation, raise awareness of evaluation methodology, and improve the use and dissemination of results to support these smaller entities.

As such, Nightingale said, her office ensured that its policy applies throughout the department and across administrations. She noted that her office drew from the work of other prominent evaluation agencies when creating its policy and takes pride in the fact that the policy has been accepted and supported throughout the department. DOL has also created an evidence-based Clearinghouse for Labor Evaluation and Research (CLEAR),[2] which contains guidelines for methodological rigor to which the agency expects both staff and contractors to adhere.

Nightingale reiterated Goldstein's point about the importance of rigor and how an evaluation policy should contain principles of rigor that apply to all types of evaluation and research. She said the focus should be on building and accumulating evaluation—that is, forming decisions that are built on a body of evidence and not just a single study—and on continuous improvement and innovation. Nightingale touched on transparency by reiterating the need to place dissemination protocol in legislation, and she closed by reminding the group that ethics should apply to both the protection of the subjects and to the integrity with which evaluations are conducted.

INSTITUTE OF EDUCATION SCIENCES

Ruth Neild (Research for Action), who previously worked at the Institute of Education Sciences [IES]), started her presentation by acknowledging that IES uses very similar strategies to those of ACF and DOL to promote transparency, rigor, independence, relevance, and ethics in its evaluations. The difference for IES, Neild explained, is that it also incorporates formal peer review of its evaluation reports to promote rigor and scientific integ-

[2] See https://clear.dol.gov [May 2017].

rity. This peer review process applies to evaluations of programs conducted by the agency and its contractors, but not to field-initiated grants.

The IES was established in 2002 as a product of the Education Sciences Reform Act (ESRA).[3] What makes IES unique, Neild pointed out, was that the ESRA charges the director of IES with ensuring that the agency's activities are "objective, secular, neutral, non-ideological, free of partisan political influence, [and] free of racial, cultural, gender, or regional bias," and it authorizes the director to publish scientific reports without approval from the Secretary or any other office—what is referred to as IES's independent publication authority (ESRA Section 186). Neild said that this authority, in addition to rigorous peer review (which is also a mandate of ESRA), makes it much more challenging for results of IES' scientific studies to be suppressed or changed to support political objectives.

Neild noted that the current evaluation budget at IES is approximately $40 million. In response to ESRA, the National Board for Education Sciences, IES' advisory board, established a Standards and Review Office (SRO) to manage the peer review process. IES evaluation staff work with program offices to identify needs for evaluation and then conceptualize what those evaluations will look like. The board provides direction to the contractors who are carrying out the evaluations, review draft reports, and determine when a report is ready for external peer review. Once the report is ready, the commissioner transmits the report manuscript to the SRO. The SRO coordinates the peer review process, which works similarly to that for scholarly journals. Neild believes that the time and staff needed to conduct the peer review process are worthwhile tradeoffs to obtain a valuable product.

Neild shifted the discussion to factors that threaten the five major principles for evaluation. Some threats are external, such as suppression or manipulation of results for political purposes. Other threats can come from within, from hasty work or a desire to capture a story in a specific way that may not be exactly what the data show. She said that external peer review helps to mitigate these risks and increase the public trust in their agency's findings. Neild said she also believes that peer review incentivizes high-quality work by staff because they know that publication is not a given: it has to be earned by producing work that meets rigorous and objective standards. Peer review pushes evaluators to provide clear explanations of the purpose, the background, the methods, and the findings for a study. Lastly, Neild said she thinks that it contributes to increased overall credibility for evaluation products originating from federal agencies.

[3] See https://www2.ed.gov/policy/rschstat/leg/PL107-279.pdf [May 2017].

MILLENNIUM CHALLENGE CORPORATION

Jack Molyneaux (Millennium Challenge Corporation [MCC]) explained that MCC is a small, independent federal agency, founded in 2004, committed to reducing poverty in well-governed low-income countries through investments in sustained economic growth. It was created in response to bipartisan interest in aiding with international assistance, which was seen as ultimately beneficial to the U.S. economy. MCC's authoring legislation contains components that ensure that international investments are being used in the right way, for the right purpose, and yielding the expected results—all of which is contingent on having a credible evaluation strategy. Congress also requires that MCC's compacts—grants provided to partner countries' governments—contain specific benchmarks, strategies, and plans for annual progress updates.

MCC's Board of Directors consists of the agency's chief executive officer, four executive branch members, and four congressional appointees. The agency's evaluation policy, first proposed in 2009 and formally adopted in 2012, mirrors those of prominent evaluation agencies in many ways, but it also has key differences. One such difference is the requirement that every project, regardless of size, be expected to undergo independent evaluation. About 97 percent of MCC's projects are subjected to independent evaluations, which accounts for about 98.5 percent of the funding: the exceptions are very small studies and a few canceled projects.

To manage cost and scope of the evaluation, Molyneaux explained, the policy is structured in a way that promotes developing evaluation design in tandem with the program design. The operations staff who work with MCC's foreign counterparts to create, implement, and maintain the projects work in country teams with MCC's evaluation staff but report administratively to a separate department. In practice, MCC implements its independent evaluations by contracting reputable evaluators who are given authority over the contents of their evaluations (subject to ethical protection of respondents' confidentiality). And although there is a process in place by which staff can provide feedback if they think a factual error or a methodological problem has arisen, evaluators have editorial independence in reporting their results; they can choose to accept or reject any feedback from the project's sponsor.

Molyneaux said the goal of the evaluations is to measure attributable impact whenever feasible and when the costs are considered warranted. Because of the nature of MCC's projects (infrastructure projects such as building roads, for example), this cannot always be completed using rigorous impact evaluations. He reiterated Goldstein's point about the need to use methods that are appropriate to each program and emphasized that rigor is still at the forefront. He explained that this process is not always smooth, but it is guided by the principles of cost-efficiency: just as MCC uses economic logic and cost-benefit analysis to inform evaluation design, the agency must also ensure that the cost of the evaluation can be justified with regard to the value of the accountability and learning it is expected to yield.

Molyneaux described some early evaluations MCC conducted, including a series of evaluations of farmer training programs. When the staff pulled together a critical mass of the evaluation results for publication, they were disappointed to find that the programs had not had the desired effects. He said it is often a challenge to develop and implement successful new programs. Most of the simple interventions that are known to work have already been exploited in MCC's partner countries. The real challenge is improving upon these already exploited opportunities. Some other early problems MCC faced were due to a lack of integration between evaluation planning and program design: in one instance, a farmer training program was executed and evaluated, even though procurement issues had delayed completion of the irrigation system the farmers were trained to use until several years after the training and the evaluation were completed. Molyneaux said that MCC is forthright with its evaluation results, even when they are disappointing, and that the ensuing open dialogue has helped the agency improve evaluation and program design. He added that he is impressed with the increased due diligence he has seen in MCC's agriculture and road development sectors as a result of the sectoral evaluation reviews, and that the irrigation infrastructure sector is on a similar positive path.

4

Refining the Tricks of the Trade

This open forum session focused on two questions: What more do we need to understand about federal evaluation? What's missing from current evaluation principles and practices? Russ Whitehurst (chair, steering committee) invited workshop participants to comment on the policies and policy-making processes at their respective agencies and organizations.

Mark Shroder (Department of Housing and Urban Development [HUD]) said that HUD's policy statement on evaluation was created after having seen the success of the Department of Labor's (DOL) and other major agencies' policy documents and, as such, closely resembles several of them. He particularly mentioned the previous discussants' points on transparency and on publishing reports regardless of findings. He noted that while he agrees that every methodologically valid report (as determined by agency staff) should indeed be published, he personally does not believe that reports of evaluations that were not found to be methodologically sound should have to be released simply because they were done. Shroder referenced the Information Quality Act (IQA), under which he said there is a little-known directive to agencies not to publish findings they do not believe to be true.[1]

Whitehurst asked if this issue could be addressed on the front end

[1] Pursuant to the IQA, agencies are to maintain and disseminate data that are found to be of good quality, objectivity, utility, and integrity of information. They are also to provide administrative mechanisms allowing relevant persons to seek and obtain correction of information deemed not to align with those characteristics: see Section 515 of the Treasury and General Government Appropriations Act, 2001 [Pub. L. No. 106-554, 44 U.S.C. § 3516 note].

instead of after the fact—if an agency could determine, before money has been spent, that an evaluation will not be methodologically valid. Shroder responded that changes in a program's expectation or scope can sometimes lead to unforeseen problems. Jack Molyneaux (Millennium Challenge Corporation [MCC]) agreed that some evaluations are indeed stronger or weaker than others, but said that MCC prefers not to be the censor: it registers all evaluations, publishes all results in the agency's evaluation catalog, and often encourages peer reviewers to weigh in on methodological quality. He also added that the methodology should be appropriate to the project and not placed generically into the evaluation requirements.

Clinton Brass (Congressional Research Service)[2] commented on the pros and cons of incorporating policies into a statute, as perceived by practitioners and advocates. Although some people believe it is useful to place methods into statutes to ensure they remain part of the discussion, others believe the inclusion yields too narrow a focus. Brass gave an example of a tiered evidence initiative that narrowly defined "evidence" (for internal and external validity) as primarily coming from impact evaluations—a point of controversy in the evaluation field. Whitehurst echoed Brass' observation, noting similar issues at the Department of Education (DoED), for example, in the Elementary and Secondary Education Act and the Education Sciences Reform Act: a strongly worded congressional preference for the use of randomized controlled trials quickly became synonymous with a need to carry them out to reinforce scientific rigor. As education program evaluation has continued to mature, Whitehurst said, he has seen the language transform into wording that calls for use of the most rigorous method appropriate to the question being asked.

Judith Gueron noted two gaps in current principles and practices. One is how the policies for evaluation are written in a "one-off" way that does not promote replication, despite evidence that replication bolsters findings. The other is the need for more focus on communication and educating the general public on the importance of evaluation, beyond simply putting reports on the internet. She said that evaluation is a big investment and agencies need to build a constituency to educate the general public and government officials on the results from and the importance of evaluation.

Thomas Feucht (National Institute of Justice) asked the panelists about independence in terms of funding: When an agency places a requirement for evaluation in its policy, does that consequently yield the type of one-off evaluations to which Gueron referred? Conversely, how are other programs' evaluations funded when they do not have the same written provisions? Is independence tied to appropriation? He mentioned that evaluations of

[2] Brass reminded participants that all of his comments throughout the workshop reflect his views and not those of the Congressional Research Service.

crime prevention programs he used to manage suffered when, without appropriated provisions, the agency's core funding was stretched too thin to fully support them.

Bethanne Barnes (Washington State Institute for Public Policy), formerly of the U.S. Office of Management and Budget (OMB), noted a paper OMB wrote for the Commission on Evidence-Based Policy on the uses of evidence that discusses how funding structures can affect the development of a portfolio of evidence.[3] Nightingale briefly described DOL's funding strategy for evaluation, which includes drawing small percentages from the department's operating budget, as well as from funding programs and discretionary grants. The evaluation office prioritizes appropriation on the basis of its agencies' needs and the need to build on evidence from previous studies, and then uses that information to create an annual evaluation agenda for the department. She said there is a "professional balancing act" that is needed when discussing funding for evaluation, which requires highlighting the importance of evaluation in light of other mission-critical activities at operating agencies. Whitehurst recalled having observed a similar system at DoED and agreed that the planning for appropriations should not come solely from within the evaluation agencies; rather, it should incorporate the positions and needs of all the stakeholders with the agency.

Before the session was brought to a close, George Cave (Summit Consulting) noted that although randomized controlled trials are an improvement over prior evaluation methods, there are different types. In addition to the "thumbs up, thumbs down" trials, there are also theory-of-change trials, in which the timing and sequencing of events in a particular treatment are used to gain insight into impacts observed in a program. He suggested that both methods be given equal consideration.

[3] See: https://obamawhitehouse.archives.gov/sites/default/files/omb/mgmt-gpra/using_administrative_and_survey_data_to_build_evidence_0.pdf [May 2017].

5

Putting Principles into Practice: A Balancing Act

Judith Gueron (member, steering committee) began the discussion on advancing high-quality evaluation by listing the six key components: protecting scientific quality, producing useful results, transparency, independence, ethical standards, and funding. The keys to protecting scientific quality, she said, are having a strong evaluation team, a strong design that addresses the appropriate questions, and review procedures that protect against false claims and reinforce credibility.

Rebecca Maynard (member, steering committee) stressed the importance of agencies "taking ownership" of an evaluation and fully understanding its purpose in order to define the appropriate strategy. An evaluation may be descriptive, causal, or a measure of change, and each should be approached from a different point of view. She also noted the importance of weighing the net cost of a study against its overall effectiveness, listing this as another dimension of quality. Demetra Nightingale (Urban Institute) reiterated that the principles and practices need to allow for the flexibility to adapt guidelines and strategies as needed depending on the study.

Naomi Goldstein (Administration for Children and Families) said she was struck by the differences she heard between peer review methods of the Institute of Education Sciences (IES), which screen studies prior to release, and methods of the Millennium Challenge Corporation, which encourage extensive peer review while still promoting release of all studies. She can see value in each approach, as long as they are carried out in a climate in which the need for high quality is valued and understood. Bethanne Barnes (Washington State Institute for Public Policy) agreed that there is a place for postrelease reviews in the discussion of scientific quality, and she noted that the Office of Management and Budget's clearinghouses conduct quality reviews of both federal and nonfederal documents. If an individual study is incorporated into a larger portfolio of work, she said, a back-end review will allow for consideration of where the study fits into the bigger picture. Clinton Brass (Congressional Research Service) raised the question of whether the definition of "scientific quality" is at all relative to the nature of the policy area, the research question being asked, or the intended use of the results.

Howard Rolston (member, steering committee), referring to Gueron's comment about the importance of design, said that there is tension between performance management and evaluation: performance management typically pays less attention to design and makes causal claims without any explicit identification of a counterfactual, which is a real issue in the use of administrative data and would not be tolerated in high-quality evaluation. He also mentioned that there are new risks presented by growing access to administrative data, as well as a challenge to find complementary evaluation designs for these kinds of datasets.

Considering that practitioners and politicians are sometimes interested in more than just the bottom line, Gueron asked to what extent evaluations, methods, and reports should include interpretations that go beyond the highest standard of rigor in order to produce even more relevant and useful results—providing insight on resource allocation, areas for program improvement, and so on—while still protecting scientific quality. She suggested that studies could distinguish results that are relatively definitive from the less conclusive results that are suggested by a pattern of findings. Jean Grossman (Princeton University and MDRC) agreed that it is not entirely fair to taxpayers if evaluators only report on the findings that are irrefutable, when the often substantial data collection and analysis efforts have generated other potential insights. Furthermore, she said that in most cases evaluators would appreciate being given the leeway to explore the mechanisms behind results—as long as they can differentiate between which aspects of an evaluation are confirmatory and which are explanatory.

Russ Whitehurst (chair, steering committee) agreed with Grossman about the usefulness of supplemental analyses and interpretations but countered that some stakeholders expect answers to specific questions, and including findings on nonessential questions in a report may expose an agency to unintended political backlash. In the early stages of IES, Whitehurst said, the Secretary of Education and other officials were eager to receive scientifically based evidence and results to justify programs like No Child Left Behind, often requesting the results before the evaluation was completed or before the study had any evidence to provide. As a solution, Whitehurst said he and his staff established practice guides, which made statements about broader topics and graded the evidence in terms of strength, type, and source (such as expert panels).

Gueron then asked the participants how they might handle a report in which the findings vary across outcomes, subgroups, time periods, and program settings (although perhaps not to a point of statistical significance) or deviate greatly from the expected results: Should one exclude a unique finding from a report if it was not part of the initial design? Goldstein responded that one should proceed with caution. It is important to highlight these kinds of differences, give the necessary caveats, and move forward to new research questions.

Miron Straf (Virginia Polytechnic Institute and State University) said that the key is often in the differences in program implementation. He said that he believes agencies should encourage exploration and not constrain the evaluators by forcing them to stick too closely to an initial protocol. He asserted that moving towards such a process of continuous improvement and experimentation will allow the flexibility to really learn what works. Maynard said the best approach would be to deliver two reports: a primary report answering the impact questions and providing an explanation of the methodology and a supplemental (possibly lengthier) report on other noteworthy issues and exploratory work. In this way it would be clear to the stakeholder that the focus has shifted to a different level of evaluation or rigor of evidence.

Gueron turned to transparency. She said that clarity, full disclosure, and careful timing were all key to convincing audiences of the credibility of an evaluation and ensuring neutrality in presentation. She asked the participants to weigh in on the pressures of timing when balanced against a desire to release complete results and whether or not either of those factors could threaten future funding or lead to undue political interference. Whitehurst emphasized the need for schedules for each component—contractors, peer reviewers, professional staff, and so on—that are appropriate to the context. A project should have neither too tight nor too distant a deadline, while still allowing a cushion. He also said it is important to make evaluation data available for secondary analysis.

Rolston said that the practice of registering studies also helps to enhance transparency. Maynard agreed that registering studies and laying out standards and expectations about evaluation methods and reporting can contribute to a smoother process. Both Rolston and Maynard noted that there have been improvements in the field in this regard. Evan Mayo-Wilson (Johns Hopkins University) wondered if participants thought there could ever be a registration mechanism for health behavior and labor studies similar to that in medical trials. Maynard reported that with support from IES, the Society for Research on Educational Effectiveness is supporting development of a platform for registering causal inference studies, which is scheduled to launch in fall 2017.

Gueron next asked the group how federal agencies can reinforce independence in evaluations and protection from pressure to bias the selection of contractors or the reporting of results while also balancing their responsibility for the study and the need to gain credibility. How much flexibility should contractors have in conducting analyses, quality control, and report dissemination? Can reliance on a technical-only review protect contractors from pressure or inclination to spin the results? Mark Shroder (Department of Housing and Urban Development [HUD]) noted that for HUD, the threat of "bias" is sometimes introduced by a requirement that the agency has to pick a small business contractor over a larger company, which automatically rules out several qualified evaluators. Brass cited Eleanor Chelimsky (2008): in some circumstances there may be a tradeoff between an evaluator's independence and an agency's capacity to evaluate and learn: for example, it may be important for a mission-oriented unit to evaluate itself and take ownership of its learning agenda.

Gueron then inquired about how far the concept of independence extends. Is a contractor seen as an extension of the agency? Does it undercut contractors' credibility if they do work for an agency seen as partisan? With that in mind, how does one attract the best people to do evaluation work? Whitehurst suggested that design competitions—in which the focus is not what work will be done, but how it will be done—can address this issue. Barnes, Rolston, and Ruth Neild (Institute of Education Sciences) all agreed that independence between federal agencies and contractors should not be viewed as an either/or situation; instead, it should be seen as a relationship that has to be managed throughout each project. Neild reiterated the utility of peer review technical working groups to mitigate the risk of bias.

Barnes reminded participants that unlike larger agencies with standalone evaluation agencies or offices, smaller agencies may have to handle evaluations for their specific program area. That difference in structure has important implications for how a program manages independence and works to ensure scientific integrity. Lauren Supplee (Child Trends) commented on how difficult it was to nail down a specific standard for "evaluator independence" during a high-stakes review she coordinated: To whom does it pertain? What if a critical person has multiple roles (a funder also being a program supporter, for example)? Her team's solution was to always report the scenarios in full and allow the user to come to its own conclusion. Supplee added that these discussions of standards could also be valuable in the academic community.

In considering ethical standards, Gueron said she has learned over time that there does not need to be a tradeoff between rigor and ethics, and she noted the critique that random assignment, while rigorous, is too demanding in certain contexts. She cautioned about ruling out random assignment too quickly and said that it is critical to be able to build a defense against ethical objections that may arise. Maynard commented that if an agency is proposing to use a method other than a randomized controlled trial to answer questions about impact or effectiveness, it should have a compelling argument as to why randomization cannot or should not be used. She said that the first thing one should do is gather information about what stakeholders believe would be challenging or unethical about randomization and what they view as the preferred alternative—which Gueron noted is often the hardest issue to counter—and then to systematically address the concerns, including the evaluation threats associated with the alternative. Christina Yancey (Department of Labor) pointed out that, at times, the most rigorous method (randomized controlled trials, for example) can overlook small or hard-to-sample populations. In these instances, she believes that the ethical approach is to still study these groups to have information on them, even if the data obtained do not meet a certain scientific standard.

Shroder raised an ethical problem: although IES, the Census Bureau, and the Internal Revenue Service have special regulations safeguarding the use of their data, many evaluation agencies do not have similar protections. Most evaluation agencies are not protected. He added that since the Freedom of Information Act often takes precedence over the Privacy Act, if a federal judge does not hold that a federal agency has shown a probability that the identity of individuals will be disclosed, the information in question must be disclosed. Whitehurst agreed on the importance of this issue.

Turning to funding, Gueron asked: If obtaining adequate funding for evaluation is so critical, are there ways to implement evaluation policies and practices that guard against political pressures tied to funding? Constance Citro (Committee on National Statistics) reiterated the need for qualified staff. She explained that the financial issue often extends to hiring caps for staff. She said it might benefit smaller agencies to learn from larger agencies that have had success with their evaluation policies and practices. Goldstein said that even when interest is high, acquiring high-quality staff within the constraints of the federal hiring system can be difficult. However, she noted, mobility of federal employees becoming contractors and vice versa sometimes aids with congruency. Neild and Nightingale added that certain staff gravitate to more hands-on work: keeping those staff engaged and encouraging them (particularly those coming from academia) to continue pursuing their research once they become federal employees can help agencies strike a balance when competing with contractors or academia to hire individuals with the needed technical qualifications.

6

Making Them Stick: Institutionalizing the Principles

William Sabol (member, steering committee) discussed his experience leading a federal statistical agency (Bureau of Justice Statistics) and with helping to develop the second edition of *Principles and Practices for a Federal Statistical Agency* (National Research Council, 2001). He asserted that those principles—relevance, credibility, trust, and a strong position of independence—are very similar to those being discussed for federal program evaluation. Sabol gave examples of how that volume addressed independence, which included:

- separation of the statistical agency from the parts of the department that are responsible for policy making and for law enforcement activities;
- control over professional actions, especially the selection and appointment of qualified and professional staff;
- authority to release information without prior clearance and adherence to predetermined schedule of release; and
- the ability to control information technology systems, tied largely to protection of data.

Sabol stressed that institutionalizing principles is not a one-and-done process: he has seen situations both in and outside his former agency that show how the principles and how agency heads' capacity to uphold them and maintain independent, objective data can be challenged in many ways. As such, he said, the principles need to be continuously negotiated and renegotiated to address both new and ongoing issues. He disagreed with the point made by Jean Grossman (Princeton University and MDRC) that the Paperwork Reduction Act had been a hindrance, primarily because it gives the U.S. Office of Management and Budget (OMB) the authority to coordinate and develop the principles and policies for the 13 primary federal statistical agencies. In addition, he said, OMB's creation of the Interagency Council on Statistical Policy and the 2002 Confidential Information Protection and Statistical Efficiency Act (CIPSEA) were very important developments in terms of refining policy and promoting governmentwide data quality standards.

Bethanne Barnes (Washington State Institute for Public Policy), speaking on her former role as head of the OMB evidence team, noted that the statistical system is one of many government functions that have a formalized structure for information sharing, policy feedback, and best practices. This structure has key components, including: a council to facilitate collaboration with OMB; a designated office within OMB to set broad policy guidance; and staff to support the council's work. She mentioned how in 2014 the OMB's Office of Information and Regulatory Affairs issued Statistical Policy Directive No. 1 (also referred to as the trust directive),[1] which essentially codifies the information in *Principles and Practices for a Federal Statistical Agency*. OMB has recently begun providing consultation to several evaluation offices on their evaluation practices, based on its experience with federal statistical infrastructure and with providing guidance on evidence-based policy. Barnes attributed the success and widespread acceptance of the statistical principles to strong interagency collaboration and emphasized the importance of sharing ideas across agencies.

Barnes acknowledged that evaluation functions do not have a similar type of overarching structure, in part because evaluation has developed more slowly, and because the nature of the structures in individual agencies has been so varied. She mentioned a report from the U.S. Government Accountability Office (2013) that showed that agencies with centralized evaluation offices had broader evaluation coverage and greater use of evaluation data. The report also noted, however, that only half of the agencies reviewed had stable sources of evaluation funding.

Barnes noted that OMB has established an evidence team within the Economic Policy Division, which focuses on multiple aspects of evidence-based policy. OMB seeks to eventually become a home for federal evaluation policy, she said. It has also informally created an Interagency Council on Evaluation Policy (cochaired by workshop participant Naomi Goldstein, of the Administration for Children and Families), which exchanges infor-

[1] *Fundamental Responsibilities of Federal Statistical Agencies and Recognized Statistical Units;* available: https://www.federalregister.gov/documents/2014/12/02/2014-28326/statistical-policy-directive-no-1-fundamental-responsibilities-of-federal-statistical-agencies-and [May 2017].

mation, collaborates on areas of common interest, and provides coordinated routine feedback to OMB on issues that affect evaluation functions. She said this council could be the basis for a more formalized structure.

Barnes said a core part of OMB's work is to help agencies with developing authorizing legislation and with funding sources and levels. In 2016, OMB updated its Circular A-11 guidance document[2] to improve the definition of evaluation, to emphasize the need for a portfolio of evidence, and to introduce the concept of credible use of evidence, including intended use of evidence. The document also includes instructions for agencies to use evaluation results and establish learning agendas in their strategic planning processes. In addition, it includes instructions to continue to use those tools throughout their performance management processes, which Barnes noted is a separate process from the development of credible evidence. She said, however, that none of those documents directly reference the principles and practices being discussed today in a comprehensive framework. Outside of OMB, Barnes said that the "Holdren memo" (Holdren, 2010) is another document that provides guidance on principles and procedures integral to protecting scientific integrity and strengthening the credibility of government research. She asserted that the central theme of the Holdren memo is that the public must be able to trust the scientific process, and it reinforces this by providing recommendations for facilitating the professional development of government scientists through such activities as publishing in peer reviewed and other scholarly journals and participating in professional meetings and societies.

Sabol asked the workshop participants what external entities could do to help institutionalize the principles and what evaluation agencies themselves could do. Russ Whitehurst (chair, steering committee) commented that Congress plays a critical role and that congressional action is most easily obtained by providing OMB with the authority to oversee the process. Clinton Brass (Congressional Research Service) said that evaluation activities seem to be Balkanized both within and among agencies—evaluation versus performance management, applied research versus methods, etc.—which may be a challenge to institutionalizing principles for evaluation and should be taken into consideration.

Sabol asked Demetra Nightingale (Urban Institute) how she managed this issue in the Department of Labor (DOL). Nightingale said that the agency maintains connections and open lines of communication among staff in statistical analysis and products, policy analysis, performance management, and data analytics, noting that evaluation touches all of these areas. She reiterated Whitehurst's point about OMB's role, noting that

[2] *Preparation, Submission, and Execution of the Budget*; available: https://obamawhitehouse.archives.gov/sites/default/files/omb/assets/a11_current_year/a11_2016.pdf [May 2017].

it has encouraged conformity among offices by requiring evidence-based justifications for budget increases and clarifying that the term “statistical purposes” includes evaluation. DOL also includes a chapter on evidence in its strategic plan. Barnes said that OMB is structured similarly to DOL and added that OMB’s Circular No. A-11 gives agencies guidance on this type of collaboration; she knows that the extent to which that is similarly executed in agencies varies widely across the government.

Sabol next asked the workshop participants how agencies identify existing guidance on applying evaluation principles in order to both take advantage of what opportunities currently exist and also to provide insight for future development. Howard Rolston (member, steering committee) noted that although it can be difficult because of the Balkanization of agencies and inconsistent support for evaluation, continued vigilance by OMB and broad congressional support can help those efforts. Thomas Feucht (National Institute of Justice) mentioned that, since the ultimate goal is to institutionalize evaluation principles and the conversation has been placed in context with the statistical framework and the value of *Principles and Practices for a Federal Statistical Agency*, acquiring something similar for federal evaluation would likely require its own statute and legislation.

Lauren Supplee (Child Trends) asked Sabol and Barnes if they could see any downside to implementing a more structured system. Sabol said that while there is potential for those in leadership to exercise more or less latitude, in general he believes that CIPSEA provides an example of implementation in a manner that is quite positive, and Barnes agreed. Daryl Kade (Substance Abuse and Mental Health Services Administration) asked if the upcoming transition to a new administration presented an opportunity to pursue institutionalizing evaluation principles more formally. Sabol referenced the Committee on National Statistics’s 4-year cycle for *Principles and Practices for a Federal Statistical Agency,* seeing a potential benefit in doing the same for evaluation and providing consistency in times of staff mobility. Christopher Walsh (Department of Housing and Urban Development) suggested that the 24 agencies subject to the Chief Financial Officers Act could use the requirement to include program evaluation in their 4-year strategic plans as an opportunity to institutionalize their principles.

Sandy Davis (Bipartisan Policy Center) said he believes that evaluation will not gain traction as an ongoing part of the policy-making process unless there is congressional support for it. He said that there appears to be a lot of congressional interest in improving evaluation, on both sides of the aisle, noting the support of Speaker Paul Ryan and Senator Patty Murray for the Commission on Evidence-Based Policymaking. He added that just like the process and legislation for congressional budgets, developing a structure for evaluation may take time and will undoubtedly require changes to authorizing legislation.

Regarding ethics, Sabol asked if the workshop participants see any constraints on staff who do evaluation and scientific work. He also asked about potential conflict-of-interest issues that may arise because of partnerships between government researchers and external entities. Feucht said that because of the nature of grants, there is a wide range of relationships between programs and external entities, which can occasionally introduce confusion. Mark Shroder (Department of Housing and Urban Development) noted that some agencies do not permit professional staff to publish their research findings without approval. He opposes this practice and believes professional staff should be free to publish so long as they clarify that their opinions may not reflect those of their agencies.

7

Cheerleaders, Naysayers, Large and Small Evaluators: Fostering Support and Inclusion

Rebecca Maynard (member, steering committee) opened the session by noting that although there are several stakeholder groups that have a vested interest in developing a principles and practices document for evaluation, there are others who would not consider it such a good idea. She referred back to earlier discussion about developing a clear strategy for the evaluation, considering who may be threatened by the outcome, and the need to incorporate evaluations into policy. Maynard said she believes strongly that there should be a push to design evaluations with the expectation that the results may be positive, neutral, inconclusive, or negative and to plan for any of these outcomes. She also does not think every evaluation should be thought of as start to finish: having fluidity and built-in decision points can be beneficial.

Jon Baron (Arnold Foundation) said he believes that any guidance document being developed should take on a “less is more” approach, highlighting a few key principles and making a persuasive case for each of them. He does not think it would be beneficial to try and cover the whole landscape of evaluation in a single document, nor does he think that a long document would be read thoroughly and carefully. In addition, Baron said he believes that a technical document would be preferable to a consensus document, which might not contain the needed clarity or specificity. He gave the example of the Common Evidence Guidelines, a joint publication of the Institute of Education Sciences and the National Science Foundation (NSF) (2013), which states that “[g]enerally and when feasible, [studies] should use designs in which the treatment and comparison groups are randomly assigned.”

Baron said he thinks these guidelines could potentially serve as a starting point for an evaluation policy document. He also said that a central goal of evaluation efforts should be to grow the body of interventions that are backed by credible evidence of effectiveness. The specific goal would be to build the number of interventions shown in high-quality randomized experiments, replicated across different studies or sites, that produce sizable impacts on important life outcomes. He gave examples of health research and social policy studies that follow this paradigm, and he asserted that this approach helps nonscientific stakeholders know and accept the value of evaluation studies. Baron said he wants to see this type of visceral demonstration of the value of research in social policy and believes that it is the key to making evaluations politically sustainable. Howard Rolston (member, steering committee) added that, because the general public is often wary that facts and figures coming from government reports appear to support specific political agendas, it is important to consider evaluation and dissemination strategies that are free of bias and preserve the facts.

Baron noted that he has also been on the other side, in a way, when highly credible evaluations produced disappointing results. He quoted Manzi's *Uncontrolled* (2012, Ch. 11), saying that "innovative ideas rarely work," and mentioned two ideas on how to increase the yield of positive findings in larger evaluations. First, he suggested that prior to funding a large randomized experiment, one look for a very strong signal from prior research or evaluation literature that the intervention being evaluated could produce meaningful positive effects—promising evidence that it could be the exception, so to speak. The second tactic Baron suggested was to make a small investment up front to discover the mechanisms and look for a large effect on proximal outcomes before going forward with a major evaluation. This small step could take the form of an initial low-cost randomized controlled trial or quasi-experiment. He also echoed Maynard's suggestion about incorporating an interim decision point for short-term follow-up in a study in which one could expect early indication of long-term outcomes.

Mark Shroder (Department of Housing and Urban Development) raised the concern that the funding for the studies and the information requests are still closely controlled by Congress and the Office of Management and Budget's (OMB) Office of Information and Regulatory Affairs, respectively. Russ Whitehurst (chair, steering committee) said he believes that, because of that control, an evaluation document of the kind being discussed should not be produced by a direct stakeholder; instead, it should be written by a foundation or similar nonstakeholder, nonpolitical organization, be addressed directly to Congress, and take the form of proposed legislation. He added that he has seen growing interest in evidence-based policy from both sides of the aisle, and he believes that evaluation has the potential to gain similar bipartisan appeal.

Judith Gueron (member, steering committee) said that in her experience foundations are sometimes less concerned with exploring new learning, often taking the position that they "already know enough"; their focus is on proving the desired outcome instead of learning whether it was worthwhile. She said she believes that foundations can be useful partners to the federal government, as they can fund essential activities that the government is less likely to fund—communication and dissemination, for example—and asked Baron how to better engage them. Baron answered that some foundations believe they are helping a program simply by making a contribution but that highlighting the importance of rigorous evaluation could go a long way in terms of measuring actual progress. He also noted that it is of key importance to learn what is important to the foundations when engaging with them. Rolston added that an "inside" effort by a key player, such as OMB, could bolster the acceptance of the principles.

Sherry Glied (New York University) asked about the issue of magnitude and power for some of the smaller evaluation agencies and what to do when the program or budget is not big enough to support a desired study. Would smaller experiments be accepted in these cases? Should quasi-experimental analysis be used routinely and be supplemented by randomized controlled trials once evidence accumulates? Maynard said she believes there is a benefit to accumulating small experiments, either through sequential replications or more formalized networked studies.

Demetra Nightingale (Urban Institute) cautioned the workshop participants that any document that might be created needs to go beyond simply covering impact evaluations, social programs, and experiments in more established programs: it also needs to be applicable to the variety of agencies trying to build evaluation offices. In response to a query from Maynard about the smaller agencies that often may not have a voice in these conversations, Nightingale explained that they are represented in cross-agency evaluation groups that OMB convenes and are actively involved in discussions about funding, strategy, design, and other concepts around evaluation. Jeff Dowd (Department of Energy) echoed Maynard's concern, cautioning the participants not to forget about smaller agencies with decentralized evaluation offices and to take the time to learn about their specific challenges.

Mark Schroeder (NSF) commented on the relationship among evidence, law, and legal writing and asked to what extent lawyers can contribute to making an effective synergy between different types of evidence. He mentioned that patent lawyers in particular could prove valuable because of their knowledge of science, in addition to law. Baron reminded Schroeder that the term "evidence" has a different meaning in law, but said that he is aware of rigorous evaluation having been introduced recently into legal contexts and can see how lawyers could use it to test different approaches in the criminal justice system.

Thomas Feucht (National Institute of Justice) identified three groups that might be opposed to a principles document: those from program agencies who may challenge the notion of rigor and ascribe more to the "I tried it and it works" philosophy; practitioners who may see an investment in evaluation as detracting from direct services; and smaller agencies whose programs or target populations may be underrepresented in a push towards randomized controlled trials. Baron replied that a response to these arguments would be to focus on evaluating components of a program rather than the entire program—e.g., looking at preschool interventions as opposed to the entire Head Start program. Naomi Goldstein (Administration for Children and Families) added that some political or high-level appointees may be resistant to strengthening the independence and transparency of evaluation activities; conversely, however, she said that private-sector organizations that routinely do this type of evaluation could be supportive.

8

Key Themes and Next Possible Steps

Russ Whitehurst (chair, steering committee) drew the participants' attention back to the scope of the workshop: evaluation of federal programs intended to affect human behavior. He added that U.S. taxpayers make a decision to fund those programs with the goal of improving opportunities and reducing identified problems and that failure to use their money in a way that can contribute to those goals is a disservice to them. He referenced the opinion of Jon Baron (Arnold Foundation) about evaluations revealing the low success rates of certain programs, but he said that incremental progress is still progress.

Whitehurst said that while the evaluation principles that are currently in place are very sound, they need legislation to help give them permanence and stability. He sees the legislation taking one of two forms:

- drafting legislation on an agency-by-agency basis that supports the creation of independent research and evaluation agencies and affords them protections and statutory guidelines (such as that for the Institute of Education Sciences), or
- aligning with the Paperwork Reduction Act, creating separate legislation and giving the U.S. Office of Management and Budget (OMB) some general authority over this function, similar to that which is in place for the statistical agencies.

Whitehurst acknowledged that funding had been addressed several times throughout the workshop, telling the group that, across the board, budgets for evaluation are comparatively smaller than the money allotted to other functions in the same programs. He thinks that budgeting for evaluation would need to be included in any legislation.

Whitehurst once again mentioned OMB as a key player in the practice of implementing evaluation principles. He added that the leading agencies have a role to play as well, but he suggested that the inclusion of Congress in the specifics of implementation could ultimately do more harm than good. He reiterated the importance of peer review as a system that can hold the producers of the work responsible for its quality, and he reminded participants of OMB's prior practice of putting a process in place to rate the quality of evaluation efforts as another accountability measure. Baron added that the proper use of peer review and techniques like specifying confirmatory versus exploratory hypotheses could also be leveraged to influence similar processes in scholarly journals, whose peer review protocol is sometimes not as rigorous.

Judith Gueron (member, steering committee) reminded Whitehurst and the participants about the earlier discussion on the tension that arises between focusing on rigor and making evaluations useful. Whitehurst acknowledged that while the issue is a real one, it is subjective, and a high program failure rate can adversely affect stakeholders' opinion of success. He reinforced Baron's point about the need to evaluate program components as opposed to entire programs—especially for larger, more established operations—in an effort to mitigate that tension.

Miron Straf (Virginia Polytechnic Institute and State University) countered Whitehurst and Baron's point, saying that the message he would like to see projected to the evaluation community would be an encouragement to move away from the myopic approach of focusing on an effect size of a single intervention and to look at the social programs as part of a complex system. He suggested looking at the major drivers and also considering the use of passive or "big data" as reinforcement. He also compared social program analysis to advances in medical research and agreed with Whitehurst that it might be wise to consider how to integrate an evaluation style that focuses on continuous improvement.

Naomi Goldstein (Administration for Children and Families) brought together the discussants' summary points by saying that while peer review can be valuable, she believes it is a practice, rather than a principle, and falls under the larger umbrella of quality control—a very important principle to be considered. Whitehurst thanked participants for sharing their thoughts, and supporting the committee's view that the field of federal program evaluation is an important enterprise that will need support to continue to grow, improve, and be recognized for the important contribution it makes to the U.S. population and the government.

References

Chelimsky, E. (2008). A clash of cultures. *American Journal of Evaluation, 29*(4), 400-415. doi:10.1177/1098214008324465.

Coleman, J. (1966). *Equality of Educational Opportunity*. Washington, DC: U.S. Department of Health, Education, and Welfare.

Devaney, B., Johnson, A., Maynard, R., and Trenholm, C. (2002). *The Evaluation of Abstinence Education Programs Funded under Title V Section 510: Interim Report*. Princeton, NJ: Mathematica Policy Research.

Gueron, J.M., and Rolston, H. (2013). *Fighting for Reliable Evidence*. New York: Russell Sage Foundation.

Haskins, R., and Margolis, G. (2015). *Show Me the Evidence: Obama's Fight for Rigor and Evidence in Social Policy.* Washington, DC: Brookings Institution Press.

Holdren, J.P. (2010). *Memorandum for the Heads of Executive Departments and Agencies*. Available: https://fas.org/sgp/obama/sciinteg.pdf [May 2017].

Institute of Education Sciences and the National Science Foundation. (2013). *Common Guidelines for Education Research and Development*. Available: https://www.nsf.gov/pubs/2013/nsf13126/nsf13126.pdf?WT.mc_id=USNSF_124 [May 2017].

Manzi, J. (2012). *Uncontrolled: The Surprising Payoff of Trial-and-Error for Business, Politics, and Society*. New York: Basic Books.

National Research Council. (1992). *Principles and Practices for a Federal Statistical Agency*. Committee on National Statistics. M.E. Martin and M.L. Straf, Eds. Commission of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2001). *Principles and Practices for a Federal Statistical Agency: Second Edition*. Committee on National Statistics. M.E. Martin, M.L. Straf, and C.F. Citro, Eds. Commission of Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

U.S. Government Accountability Office. (2013). *Program Evaluation: Strategies to Facilitate Agencies' Use of Evaluation in Program Management and Policy Making*. GAO-13-570. Available: http://www.gao.gov/products/GAO-13-570 [May 2017].

Appendix A

Workshop Agenda

Principles and Practices for Federal Program Evaluation

October 27, 2016
Keck Center of the National Academies of Sciences, Engineering, and Medicine
500 Fifth Street, NW, Washington, DC
Room 100

Thursday, October 27, 2016
Open Session, 9:00am–4:30pm

9:00am Call to Order
Breakfast available outside the meeting room

Welcome and Introduction
Connie Citro, *Director, Committee on National Statistics*
Christine Fortunato, *Administration for Children and Families*

Purpose of the Workshop
Russ Whitehurst, Brookings Institution, *Committee Chair*

The steering committee for the workshop will facilitate discussion on principles and practices for federal program evaluation, to include reviews of extant policies issued by the Administration for Children and Families, the Institute for Education Sciences, the Chief Evaluation Office in the Department of Labor, and other federal agencies. Throughout the workshop we will consider ways to build upon these documents, including ways to institutionalize the principles, with the goal of bolstering the integrity and protecting the objectivity of the evaluation function in federal agencies, which is essential for evidence-based policy making.

Scope of the Workshop: *Evaluations of interventions, programs, and practices intended to affect human behavior, carried out by the federal government or its contractual agents, domestic and abroad, and leading to public reports sponsored by the federal government that are intended to provide information on their impacts, cost, and implementation.*

9:20am **History of Federal Program Evaluation**
Moderator: Howard Rolston, Abt Associates (Committee Member)

Discussants: Jean Grossman, Princeton University; Ron Haskins, Brookings Institution; Larry Orr, Johns Hopkins University

History of federal program evaluation and its successes, challenges, and vicissitudes from a variety of perspectives, including formal federal government evaluation leaders, evidence-based policy advocates, and social policy researchers and producers.

10:20am Break

10:35am **Review of Present Principles by Topic and Agency Coverage**
Moderator: Brian Harris-Kojetin, CNSTAT Deputy Director

Discussants: Naomi Goldstein Administration for Children and Families (ACF); Jack Molyneaux, Millenium Challenge Corporation (MCC); Ruth Neild, Institute of Education Sciences (IES); Demetra Nightingale, Department of Labor (DOL)

Exploration of several prominent evaluation agencies and their approaches to protecting the integrity and objectivity of evaluation work through formal guidance and other means.

11:30am **What More Do We Need to Understand/What's Missing from Current Evaluation Principles & Practices: A Discussion**
Moderator: Russ Whitehurst, Brookings Institution (Committee Chair)

(For discussion panels, each moderator will open the session with initial remarks that will lay out a coherent framework for the task and issues at hand. The steering committee members will offer their opinions, and then attendees will have an opportunity to share their thoughts and insight. We encourage active audience participation during these sections.)

12:00pm Lunch

1:00pm **Issues and Challenges for Implementing Principles & Practices: A Discussion**
Moderator: Judy Gueron, MDRC President Emerita (Committee Member)

Although the five principles in current agency guidance—rigor, relevance, transparency, independence, and ethics—seem uncontestable, challenges arise in balancing them, especially since advancing high-quality evaluations requires both obtaining sustained funding and engaging the best talent. This discussion will focus on the components necessary to advance high-quality evaluations, protect the infrastructure that supports them, and ensure that the evaluations produce results that a broad community of politicians, practitioners, and funders consider objective and useful.

2:00pm **How Do We Institutionalize the Major Principles? (Discussion)**
Moderator: Bill Sabol, Westat (Committee Member)

Discussant: Bethanne Barnes (OMB)

Even with broad support for enhancing the principles outlined in the previous discussion—quality, utility, transparency, independence, and high ethical standards—these goals cannot be achieved by the efforts of evaluation staff alone. Rather, these principles need to be institutionalized into agency practices so that evaluators are protected against efforts to quash them. Similarly, ensuring an adequate funding base for evaluations requires agency-level commitments to developing knowledge both about what works and the circumstances under which something works or does not work. Institutionalizing these principles may require legislative changes to statutory authorities, new organizational entities, or new *relationships between organizations. This discussion will focus on approaches that can be taken to institutionalize quality, relevance, and independence, the opportunities and challenges associated with various approaches, and pathways and priorities to implement the changes.*

2:45pm Break

3:00pm **Garnering Support and Maintaining Focus: A Discussion**
Moderator: Rebecca Maynard, University of Pennsylvania (Committee Member)

Discussant: Jon Baron, Arnold Foundation

Discussion will center on how to develop guidance that will directly serve the interests of evaluation offices within federal agencies while simultaneously mitigating the potential resistance from offices and organizations whose interests may be threatened by the formation of such a document (e.g., advocacy organizations and special interest groups). The discussion will consider vehicles and resources that can maximize support for objective evaluation across federal, state, and local levels of government, including executive and legislative branches.

4:00pm **Future of Principles & Practices for Federal Program Evaluation: A Discussion about Next Steps**
Moderator: Russ Whitehurst, Brookings Institution (Committee Chair)

This concluding session will focus on themes from the preceding sessions and consider potential future steps for articulating and strengthening principles and practices for federal program evaluation and the evaluation function itself.

4:30pm Adjourn

Appendix B

Biographical Sketches of Steering Committee Members and Speakers

BETHANNE BARNES ***(Speaker)*** is director of the Washington State Institute for Public Policy. Previously she served as special advisor for evidence-based policy at the U.S. Office of Management and Budget (OMB). As head of OMB's evidence team, her work focused on helping federal agencies strengthen their capacity to use and build evidence to improve their effectiveness. She also worked on a variety of job training and social safety net programs at OMB, as well as cross-agency data access and evidence-building policy issues. She has a bachelor's degree from Evergreen State College and a master's degree in public administration from the Evans School of Public Affairs at the University of Washington.

JON BARON ***(Speaker)*** is vice president of evidence-based policy at the John and Laura Arnold Foundation, responsible for the foundation's strategic investments in rigorous research of evidence-based social programs and scaling those shown to produce meaningful improvements in people's lives. Previously, he founded and served as president of the Coalition for Evidence-Based Policy, a nonprofit, nonpartisan organization that worked to advance important evidence-based reforms. He previously served as a presidentially appointed member and chair of the National Board for Education Sciences and as counsel to the Committee on Small Business of the U.S. House of Representatives. He is a fellow of the National Academy of Public Administration, an honorary fellow of the Academy of Experimental Criminology, and a recipient of the Public Service Award of the Society for Prevention Research. He has a B.A. from Rice University, a master's degree in public affairs from Princeton University, and a law degree from Yale Law School.

NAOMI GOLDSTEIN ***(Speaker)*** is deputy assistant secretary for the Office of Planning, Research and Evaluation (OPRE) at the Administration for Children and Families at the U.S. Department of Health and Human Services (HHS), where she earlier served as director of OPRE's Division of Child and Family Development. Previously, she directed the United States Postal Service Commission on a Safe and Secure Workplace, an independent commission that examined workplace violence affecting the postal service and the nation. She also previously served as project manager at the Urban Institute and as executive officer in the Office of the Assistant Secretary for Planning and Evaluation at HHS. She was awarded the presidential rank of distinguished executive. She has a B.A. in philosophy from Yale University, a master's in public policy from the Kennedy School of Government at Harvard University, and a Ph.D. in public policy from Harvard University.

JEAN GROSSMAN ***(Speaker)*** is on the faculty of Princeton University's Woodrow Wilson School of Public and International Affairs and a senior research fellow at MDRC. Previously, she held positions at Public/Private Ventures and Mathematica Policy Research and served as the chief evaluation officer for the U.S. Department of Labor overseeing all of the department's program evaluations. Her work focuses on programs that serve disadvantaged youth, especially mentoring programs and out-of-school time programs, as well as on the mechanisms of mentoring, exploring the role of the match length, rematching, and the quality of the relationship. She has a Ph.D. in economics from the Massachusetts Institute of Technology.

JUDITH GUERON (***Member, Steering Committee)*** is an independent scholar in residence and president emerita at MDRC, a nonprofit organization involved in designing interventions, evaluating programs, and providing technical assistance for social programs using strict research standards. At MDRC, she directed many of the largest federal and state evaluations ever undertaken of interventions for low-income adults, young people, and families. She is a past president of the Association for Public Policy Analysis and Management (APPAM), has served on several federal advisory panels, and has frequently testified before Congress. She is a member of the board of directors of Alcoa and of the National Bureau of Economic Research and the Society for Research on Educational Effectiveness. She is a recipient of the Myrdal Prize for Evaluation Practice of the American Evaluation Association, the inaugural Richard E. Neustadt Award from the John F. Kennedy School of Government at Harvard University, and APPAM's Peter H. Rossi Award for contributions to the theory or practice of program evaluation.

She has a B.A. summa cum laude from Radcliffe College and a Ph.D. in economics from Harvard University.

RON HASKINS *(Speaker)* is a senior fellow in the Economic Studies Program and codirector of the Center on Children and Families at the Brookings Institution and senior consultant at the Annie E. Casey Foundation. Previously, he served as the senior advisor to the President for welfare policy and as a member and director of the staff of the Human Resources Subcommittee of the Ways and Means Committee of the U.S. House of Representatives. Prior to his government service, he was a senior researcher at the Frank Porter Graham Child Development Center at the University of North Carolina (UNC), Chapel Hill, a lecturer on history and education at UNC, Charlotte, and a lecturer in developmental psychology at Duke University. His areas of expertise include welfare reform, child care, child support, marriage, child protection, and budget and deficit issues. He has a bachelor's degree in history, a master's degree in education, and a Ph.D. in developmental psychology, all from the University of North Carolina, Chapel Hill.

REBECCA MAYNARD (***Member, Steering Committee)*** is on the faculty of the Graduate School of Education at the University of Pennsylvania, where she previously directed the university's predoctoral training program in interdisciplinary methods for field-based education research. Previously, she served as commissioner of the National Center for Education Evaluation and Regional Assistance at the Institute of Education Sciences at the U.S. Department of Education, where she oversaw the institute's evaluation initiatives, the What Works Clearinghouse, the Regional Education Laboratories, and the National Library of Education (including ERIC). She also previously served as senior vice president at Mathematica Policy Research, Inc. Her work focuses on the design and conduct of randomized controlled trials in the areas of education and social policy. She has a Ph.D. in economics from the University of Wisconsin–Madison.

JACK MOLYNEAUX *(Speaker)* is director of independent evaluations at the Millennium Challenge Corporation, where he also works as an applied microeconomist Previously he implemented and managed impact evaluations for Indonesia, working with the Rockefeller Foundation, the University of Indonesia, Statistics Indonesia (the Indonesian statistical agency), the RAND Corporation, and the World Bank. At the World Bank, he coordinated impact evaluations of sanitation and hygiene investments in five countries. His work focuses on evaluation and analysis in the fields of health, reproduction, nutrition, water, sanitation and hygiene, labor, education, agriculture, transportation and prices and wages. He has a Ph.D. from the University of North Carolina, Chapel Hill.

MARTHA MOOREHOUSE (*Member, Steering Committee)* is an independent consultant in Los Altos, CA. She was formerly director of the education program at the Heising-Simons Foundation, which focused on children from birth to 8. Previously, she served as senior advisor for evaluation policy for human services at the Office of the Assistant Secretary for Planning and Evaluation (ASPE) at the U.S. Department of Health and Human Services, where she also served as the director of ASPE Children and Youth Policy Division. She also previously served with the evidence team at the U.S. Office of Management and Budget and as was on the psychology faculty at the University of California, Santa Cruz. Her worked has focused on research, practice, and policy concerning children and their families. She has a Ph.D. in developmental psychology from Cornell University.

RUTH NEILD *(Speaker)* is director of research for Action's Philadelphia Education Research Consortium. Previously she was deputy director for policy and research at the Institute of Education Sciences (IES) at the U.S. Department of Education. Prior to being deputy director she served as commissioner of the National Center for Education Evaluation and Regional Assistance. Her work at IES included reorienting the federal regional educational laboratories network toward research-practice partnerships, increased the reach of the What Works Clearinghouse through improvements in dissemination and communication, and oversaw federal evaluations. Prior to her government service, she was a research scientist at the Johns Hopkins University Center for Social Organization of Schools, where she worked on research projects that ranged from descriptive and correlational to studies of impact. She has an A.B. in history and sociology summa cum laude from Bryn Mawr College and a Ph.D. in sociology from the University of Pennsylvania.

DEMETRA NIGHTINGALE *(Speaker)* is an institute fellow at the Urban Institute, where her research focuses on social, economic, and labor policy issues. She was chief evaluation officer for the U.S. Department of Labor from 2011 to 2016, responsible for coordinating the department's evaluation agenda and working with all its agencies to design and implement evaluations. She is an expert in employment policy, workforce development, labor markets, and social policies and programs, and has conducted many evaluations of federal, state, and local programs aimed at increasing employment, skills, and income for workers and families. She also teaches program evaluation at the Trachtenberg School of Public Policy and Public Administration at George Washington University. Previously, she was a senior fellow at the Urban Institute and served on the faculty at Johns Hopkins University's graduate program in public policy. She has been a senior research consultant with the World Bank and was an expert advisor to President's Clinton Welfare Reform Working Group. She has a B.A. in political science and Ph.D. in public policy, both from George Washington University.

LARRY ORR *(Speaker)* is an associate at the Institute for Policy Studies at the Bloomberg School of Public Health at Johns Hopkins University, where he teaches program evaluation. He also works as an independent consultant on the design and analysis of evaluations of public programs, and he is currently serving as an evaluation specialist on an evaluation of results-based aid in the education sector in Ethiopia for the U.K. Department for International Development. Previously, he worked at Abt Associates and in the U.S. government, holding positions at the Office of Economic Opportunity, U.S. Department of Health, Education, and Welfare, and the U.S. Department of Labor. His work involved responsibility for the design and oversight of a number of large-scale surveys and field studies, including the Panel Study of Income Dynamics and the National Job Training Partnership Act Study. He has a Ph.D. in economics from the Massachusetts Institute of Technology.

HOWARD ROLSTON (*Member, Steering Committee)* is an independent consultant in Arlington, VA. He was previously a principal associate at Abt Associates, where he directed two large-scale, multi-site random assignment evaluations: the pathways for advancing careers and education project for the Administration for Children and Families and the benefit offset national demonstration for the Social Security Administration. Prior to Abt Associates he served at the U.S. Department of Health and Human Services, where he worked in evaluating welfare reforms, employment programs, early childhood interventions, and other social programs. He has a Ph.D. in philosophy from Harvard University.

WILLIAM SABOL (*Member, Steering Committee)* is a vice president at Westat, overseeing projects on justice, child welfare, and family services. Previously, he was the director of the Bureau of Justice Statistics at the U.S. Department of Justice (DOJ), where he was responsible for managing data collection and statistical operations, developing administrative records, publishing statistical reports, and coordinating and implementing comprehensive statistical program plans. He also previously served as acting director of the DOJ's National Institute of Justice, as assistant director for homeland security and justice at the U.S. Government Accountability Office, as associate director of the Center on Urban Poverty and Social

Change at Case Western Reserve University, and as senior research associate at the Urban Institute. He has a Ph.D. in policy research and analysis from the University of Pittsburgh.

GROVER WHITEHURST (***Chair, Steering Committee***) is a senior fellow in the Center on Children and Families in the Economic Studies program at the Brookings Institution. Previously, he was the first director of the Institute of Education Sciences at the U.S. Department of Education, which received a citation from the Office of Management and Budget for having "transformed the quality and rigor of education research within the Department of Education and increased the demand for scientifically based evidence of effectiveness in the education field as a whole." He also served previously as chair of the Department of Psychology at the State University of New York at Stony Brook and academic vice president of the Merrill-Palmer Institute. His specializations include program evaluation, teacher quality, preschools, national and international student assessments, reading instruction, education technology, and education data systems. He has a Ph.D. in experimental child psychology from the University of Illinois at Urbana-Champaign.

COMMITTEE ON NATIONAL STATISTICS

The Committee on National Statistics was established in 1972 at the National Academies of Sciences, Engineering, and Medicine to improve the statistical methods and information on which public policy decisions are based. The committee carries out studies, workshops, and other activities to foster better measures and fuller understanding of the economy, the environment, public health, crime, education, immigration, poverty, welfare, and other public policy issues. It also evaluates ongoing statistical programs and tracks the statistical policy and coordinating activities of the federal government, serving a unique role at the intersection of statistics and public policy. The committee's work is supported by a consortium of federal agencies through a National Science Foundation grant, a National Agricultural Statistics Service cooperative agreement, and several individual contracts.