UNITED STATES EVALUATION POLICY:
A THEORETICAL TAXONOMY

A Dissertation
Presented to the Faculty of the Graduate School
of Cornell University
In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Margaret Alice Johnson

January 2012

UNITED STATES EVALUATION POLICY:
A THEORETICAL TAXONOMY

Margaret Alice Johnson, PhD
Cornell University 2012

Efforts are currently underway in the US federal context to improve and strengthen evaluation practice and increase the use of evaluation results to inform policies and programs. However, these efforts remain unrealized, due partly to the lack of a comprehensive theoretical framework that views evaluation and related organizational processes and institutions as part of a larger system. Early intuitive theoretical taxonomies of evaluation policy suffer from the lack of connection to specific examples and instances, and are missing clear classification criteria that would allow practical application. To generate a grounded taxonomy of evaluation policy, this study surveyed members of the American Evaluation Association in 2009, asking them to generate examples of evaluation policy, and then to sort and rate these suggested policies. Results are analyzed using the concept mapping method of Trochim (1989), which first translates aggregate sorting decisions into conceptual "distances" on a two-dimensional dot map, then uses hierarchical cluster analysis to generate groupings of ideas. These groupings become the foundation for categories in a theoretical taxonomy. Findings reveal several different dimensions by which participants grouped evaluation policies, including the dimensions of "value" and "policy mechanism." A values-by-mechanisms taxonomy and instructions for its use in an evaluation policy inventory process are proposed.

BIOGRAPHICAL SKETCH

Margaret Alice Johnson was born in Bangor, Maine and grew up in Potsdam, New York, the daughter of Dr. Arthur L. Johnson, professor of Canadian and US history at SUNY Potsdam, and Anne H. Johnson, musician, gourmet baker, masseuse and adoption finder. Her younger sister, Laura Johnson Whalen, is now a counselor and trainer of other counselors in the New York State correctional system. Margaret's growing up years were graced by the deep beauty of the Adirondacks, the richness of many musical and theatrical productions at the Crane School of Music, and the company of good friends.

She completed her Bachelor of Arts degree in political science and French at SUNY Potsdam. Her undergraduate academic advisor was John K. White, now at Catholic University. As part of her undergraduate experience, she was privileged to sing in the Crane Chorus under the late, great conductor Broch McElheran of Canada. Also as an undergraduate, she spent an academic year at l'Université de Grenoble in France, where she learned to ski and to love good wine and cheese. She then completed a semester-long internship in the New York State Assembly in the Office of Assemblywoman Cynthia Jenkins of Queens.

Upon graduation, she came to Ithaca in 1987 to work in the Cornell University Office of Government Affairs under Stephen P. Johnson, and then served as legislative aide to Assemblyman Marty Luster from 1989 until his retirement in 2002. During this time she met and married her husband, Michael Roman, and welcomed a son, Nathan and daughter, Rachael. She worked briefly for the Ithaca Displaced Homemakers' Center under the incomparable Dammi

Herath. Then in 2003, inspired by independent evaluator Marilyn Ray of the Finger Lakes Center for Law and Social Policy and encouraged by Michael Koplinka-Loehr, then a career developer for the Cornell Institute for Public Affairs (CIPA), she returned to school in 2003 to complete her MPA at CIPA with a focus in social policy. She completed her degree under the guidance of thesis advisor Professor John Bishop of the Cornell School of Industrial and Labor Relations. She entered the PhD program in the Department of Policy Analysis and Management in the School of Human Ecology at Cornell University in 2005. During her years in the PhD program, she has been very fortunate to work with Professor Bill Trochim's research and facilitation team in the Cornell Office for Research on Evaluation (C.O.R.E.).

She looks forward with great eagerness to the next interesting project.

DEDICATION

This work is dedicated, with love and gratitude

to my beautiful grandmother, Alice Hansen Hastings.

ACKNOWLEDGMENTS

The process of getting a PhD made me feel alone at times, but this was an illusion, and one I am happy to shed. First, I want to thank my Mom and Dad for giving me such a good start, exposing me to all the finer things in life, including intellectual pursuits. I want to thank my husband, Michael Roman for his patient love, steady optimism and unfailing support, both emotional and technical, from the moment he encouraged me to enter this degree program in 2005 to the final edits of my dissertation late in 2011. He has had my back in ways too numerous to mention. I also thank my children, Nathan and Rachael, for challenging me to explain the meaning and value of my work, and for always believing in me.

My gratitude goes to David Faulkner, PhD for so ably coaching me to a successful GRE after 17 years away from school. Thank you to the Cornell Institute for Public Affairs for welcoming me to Cornell and giving me the opportunity to tool up in math and economics. I'm very grateful also to Professors Jerry Ziegler, Elizabeth Peters, and John Cawley who wrote letters of recommendation to help me get into the Policy Analysis and Management PhD program.

I want to thank my committee chair, Professor Bill Trochim, for taking me on as an advisee at a time when there were few professors left in evaluation at Cornell. Without his support, it would have been truly impossible for me to complete this degree in this place. I owe him my deepest gratitude for connecting me with the larger world of evaluation thought and for opening the door to meaningful work and lively colleagues at C.O.R.E.. Huge thanks go to Professor Maureen Waller for her consistently thoughtful and thorough critique of my work, and for her patience and compassion in taking the time to talk with me through the ups and downs.

A big thank you goes to Professor Dan Lichter for challenging me to situate my work in the larger scholarly context.

Heartfelt thanks go to Marilyn Ray, PhD for providing not only the original inspiration to become an evaluator, but also ongoing encouragement and numerous opportunities for practical evaluation work outside of academia—both much needed.

Monica Hargraves, PhD deserves my deepest gratitude for her boundless willingness to talk about my research interests, and for repeatedly applying the oxygen mask of confidence throughout this process (see Shakespeare's Sonnet 29). She also deserves special acknowledgement for helping me think through the instruments for the concept mapping data collection in this study, and for her very helpful input into the naming of clusters.

I am also very grateful to Claire Hebbard, for her steady, no-nonsense perspective and for her deft handling of tricky logistics at crucial times. Merci to Tom Archibald for so generously inviting me back to share part of his office space at C.O.R.E., and for countless rich and sustaining conversations about our respective projects.

Thank you to the Cornell Graduate School for providing a facilitated dissertation support group, to the Cornell Knight Institute for Writing and to Keith Hjortshoj's for his incredible little publication, "Writing from A to B" (2010)—hurray! Thanks also to my fellow dissertators Hannah, Suzanne and Jisung, and to Maggie and Delphia in the Policy Analysis and

Management program with me, for letting me know I was not alone, and for the priceless opportunity to travel with them as they journey forward down this road.

Thank you from the heart to my best girlfriends from Potsdam growing up years, Eileen the musician and Abra the poet, for their constant affection and fierce loyalty in dark times, and for wholeheartedly celebrating small milestones. Next year in the Netherlands!

Last but not least, I want to extend a warm thank you to my friends at the Finger Lakes Cycling club, at the Ride for Life, and to all my cycling friends, especially “Flame”, for sharing those timeless moments of beauty, fresh air and freedom that have kept me sane and happy these past few years.

TABLE OF CONTENTS

LIST OF APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

> *"Although the Federal government has long invested in evaluations, many important programs have never been formally evaluated, and the evaluations that have been done have not sufficiently shaped Federal budget priorities or agency management practices."*
> ~Peter Orzag, Director of the United States Office of Management and Budget (October 2009)

> *"Finding a balance between 'getting the work done' on one hand and capacity building, on the other, may be difficult."*
> ~Clinton Brass, Congressional Research Service (January 2011)

The field of evaluation was born in the United States, yet we lag behind other nations in developing a strong evaluation system in government. Efforts are currently underway, both within and outside the government, to improve and strengthen evaluation activities and the use of evaluation results towards better policies and programs. However, these efforts remain unrealized, due partly to the lack of a comprehensive approach that views evaluation and related organizational processes and institutions as part of a larger evaluation system. Toward that end, theorists have begun to articulate the scope and boundaries of the concept of evaluation policy, and describe its relevant sub-types in the U.S. federal context (see Mark, Datta and Trochim 2009). This work could become the foundation for a comprehensive inventory of federal evaluation policies towards reform of the U.S. evaluation system. However, previous theoretical formulations suffer from the lack of connection to specific examples and instances (Blumer 1954) and a lack of classification criteria that might render them useable by organizations developing their evaluation policies, or by researchers in theory building and testing. To refine and operationalize the construct of evaluation policy, this study uses a concept mapping approach with cluster analysis to elicit examples of evaluation policy from evaluation practitioners and researchers. It then groups their ideas to build categories towards a taxonomy with workable classification criteria.

Background

What are evaluation policies, and why do they matter? Evaluation policies govern the way evaluation is conducted and the use of evaluation results. Evaluation policies convey an organization's intentions to be accountable to constituents, to use evidence in decision making, and to encourage organizational learning. Evaluation policies have the potential to influence decision making in a myriad of ways. Policies about when to evaluate outcomes, whether early or late in the program's development, can help determine whether the program is able to demonstrate success. Policies about what constitutes "success" can determine whether a program or policy is found to be successful. Policies about how to evaluate--which methods and approaches--can make the difference between a finding of "effect" or "no effect". Policies about who will conduct evaluations, whether external professionals or internal program administrators, and where in the organization internal evaluators are situated, can affect whether evaluation results are used. Policies that link evaluation results to funding can increase accountability or, alternatively, introduce incentives to game the system. Policies about how much influence to give evaluation results, whether little or much, can determine which programs policymakers choose to continue, improve or terminate (Chelimsky 2009, Datta 2009, Mark 2009, Trochim 2009).

How important is evaluation policy, in the larger scheme of things? If we stand back and regard evaluation policy in the context of larger systems, it is just one of the factors that help determine evaluation practices. Evaluation practices are, themselves, just one of the factors that help influence decision making about policies and programs. Formal decisions about policies and programs are just one of the factors that help shape the day-to-day practices of policy and program implementers, which are, in turn, just one factor affecting the quality of life of the intended targets of policies and programs (Weiss, Graham and Birkeland 2005).

Yet recent history shows that evaluation policies can have important effects. In early 2003, the U.S. Department of Education touched off a firestorm by proposing to give priority in funding to programs using the experimental, randomized controlled trial (RCT) design in their evaluations (see U.S. Department of Education, "Scientifically Based Evaluation Methods," 70 *Federal Register* 3586, Jan. 25, 2005). Nearly 300 parties responded during the month-long comment period, largely in opposition. According to an account by the Congressional Research Service, "Many critics of the Education Department priority argued that RCTs have been oversold in terms of their practical capabilities to contribute to understanding of causes, effects, impacts, "effectiveness", and in some cases, to making claims of causal relationships (even if some of the designs are not intended to calculate impacts); and that the ED priority would detrimentally affect overall priorities for education." (United States Congressional Research Service 2006, p. 24)

Nevertheless, in 2004, the United States Office of Management and Budget released a guidance documented called "What Constitutes Strong Evidence of a Program's Effectiveness" that highlights RCTs as the best way to establish evidence of effectiveness (United States Office of Management and Budget, 2004) Later, in the midst of the campaign that would elect President Obama, the American Evaluation Association responded to the memo by calling for policies to instead offer guidance on how to identify appropriate methods from among a larger spectrum of rigorous methods, and to acknowledge the value of mixing RCTs with other methods (American Evaluation Association, 2008).

Despite vocal opposition to the proposed priority and a statement by the Office of Management and Budget recognizing that RCTs are not appropriate for all programs, the priority was approved in early 2005, largely unchanged. Shortly thereafter, in 2006, President Bush used the lack of RCT evidence to justify cutting dozens of programs (Congressional Research Service 2011).

The RCT controversy caused professional evaluators to openly question whether policy on evaluation should exist at all, and if so, what form it should take. Julnes and Rog (2007) argue that evaluation policies should exist, since, in economic terms, causal information is a public good, and there are times when the "market" may not produce enough without policy intervention. However, they argue against policies that mandate particular methods, suggesting instead more flexible guidelines laying out contingencies, indicating which methods to use in which settings. They suggest that government policies should support a mixed portfolio of studies with a balance of types, including continuous improvement studies, performance (or outcome) studies, implementation studies and field trials.

Choice of evaluation method is just one area of contention. Evaluators often face pressure from stakeholders in the choice of questions they ask, the measures they use, and the way they report results (Chelimsky 2009). Especially for evaluators who take a participatory approach, convictions about rigor and quality in evaluation practice stand in tension with the need to incorporate the viewpoints of multiple evaluation stakeholders, some of whom may be significantly affected by the results. Evaluation policy in the United States has been developed through a variety of processes, some carefully considered and open, some rushed, reactive and closed (Mark, Cooksy and Trochim 2009). A more proactively and systematically developed set of evaluation policies could help support and maintain a healthy evaluation system within U.S. government (Trochim 2009, Datta 2009)

The literature on evaluation systems suggests such systems do not spontaneously arise, but require careful and conscious design. Wholey (1970) suggests each United States federal agency should have a clear definition of program objectives and output measures and the development of evaluation work plans as part of an overall strategy that prioritizes evaluation questions. Leeuw and Furubo (2008) argue that all effective evaluation systems have a clearly stated epistemological perspective, and established organizational responsibility for evaluation and its permanence. Dahler-Larsen (2006) lists among the necessary ingredients for construction of a working "evaluative information system" an evaluation unit situated in such a way as to command a critical mass of human resources, as well as managerial attention and legitimacy, since these will determine the design and content of the system, its chances for successful implementation, and its survival over time (p. 70). Other key ingredients include clearly stated evaluation criteria for quality, a self-representation that explains the justification for the evaluation system, and sufficient financial and political, support, as including the full buy-in of those implementing and collecting data within the system. According to Dahler-Larsen, all these factors must somehow be brought into alignment with each other for an information system to adequately build and sustain a healthy evaluation function. Establishing a comprehensive set of mutually complementary evaluation policies is one way to accomplish this, especially if policies are developed through a process that is open and inclusive of stakeholders (Mark 2009).

Context of the study

Early in 2009, responding to an invitation from the Obama administration, the American Evaluation Association issued a statement of principles entitled “An Evaluation Roadmap for a More Effective Government” calling for the integration of program evaluation as an essential management function of government (AEA 2009). President Bush had signed, a little over a year before, Executive Order #13450. E.O. #13450, which centralized in the Office of Management and Budget and the White House decision making about agencies’ Government Performance and Results Act compliance efforts, required all agency heads to designate Performance Improvement Officers (PIOs), and established a Performance Improvement Council (PIC), composed of PIOs, to be coordinated by the Office of Management and Budget. While the focus of Executive Order 13450 is mainly on performance measurement, the AEA Roadmap offers a set of principles for developing agency-level policy on evaluation, including: 1) scope and coverage of evaluation; 2) management of evaluation; 3) protecting quality and independence of evaluation and 4) transparency in goal setting, evaluation methods and results. The Roadmap also suggests various ways of organizing the evaluation function within an agency, emphasizing that different agencies’ evaluation needs vary depending partly on the structure of their programs, so they should be free to shape their own evaluation policies. This statement of principles calls for a “government-wide effort” and suggests that the agencies themselves develop written evaluation policies across and within federal agencies.

Shortly after the release of AEA's "Roadmap", Trochim (2009) published an intuitive taxonomy of evaluation policy, with eight evaluation policy types depicted as the slices in a layer cake. The eight policy type "slices" in this taxonomy are Goals, Participation, Capacity Building, Management, Roles, Process and Methods, Use and Meta-evaluation The cake is also divided into rings, with general policies in the outer rings and related, but progressively more specific sub-policies falling in the rings progressively closer and closer to the center. Following the work of Carver (2006) on board governance structures, for hierarchical organizations, Trochim assigns each layer of the cake a level of the organizational hierarchy. In hierarchical organizations, he argues, general policy (such as a statement of general principles) is best set at the top, with successively more specific policies delegated to successively lower levels of the organization. At the lowest level are specific step-by-step procedures (see Figure 1 below).

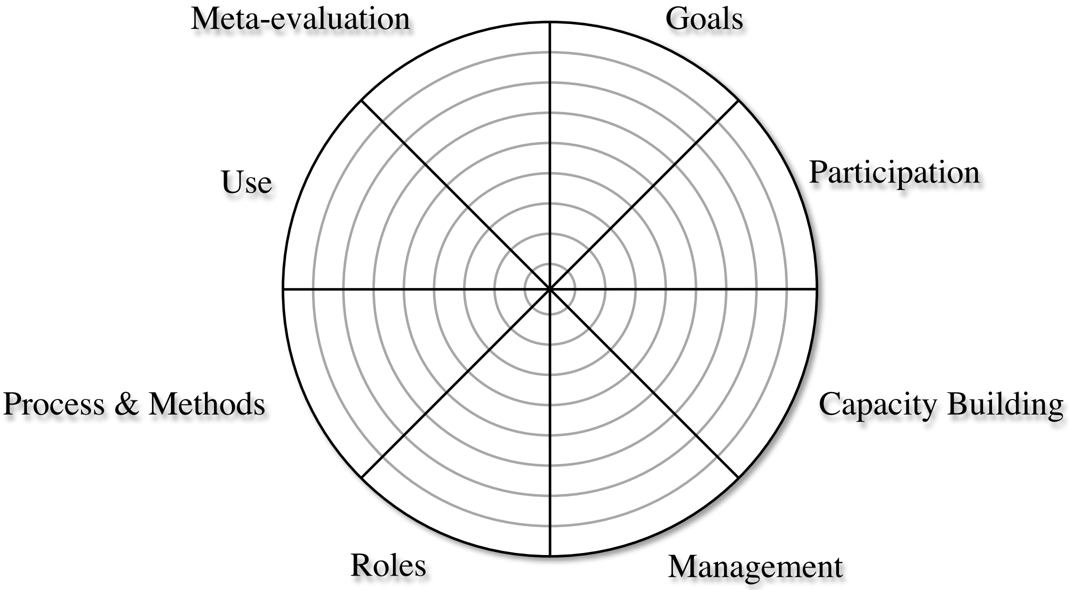

Figure 1: "The Policy Wheel"[1]

[1] Published in Trochim, W.M.K. (2009). Evaluation policy and evaluation practice. In W.M.K. Trochim, M. M. Mark, & L. J. Cooksy (Eds.), Evaluation policy and evaluation practice. New Directions for Evaluation, Issue 123, pp. 13–32, used with permission of the author.

Cooksy (2009) and Mark (2009) call for wider input by evaluation practitioners and researchers in further developing and refining the concept of evaluation policy and Trochim's taxonomy. They argue this is needed in order to identify any missing, high-level categories, to assure a comparable level of generality for all categories, and more clearly establish the boundaries of the conceptual domain (Cooksy 2009).

The data for the present study was gathered in May of 2009, seven months after a November 2008 speech to the AEA by Trochim unveiling his "policy wheel" taxonomy, and a few months after the AEA Roadmap was released, at a time when rank and file AEA members had not yet been surveyed on evaluation policy. Those who received my survey invitation for this study may have been especially interested in responding, since for many, this was their first opportunity to voice their views on evaluation policy. It seems likely that many of those who responded to my study had heard Trochim's November 2008 speech or had read the AEA Roadmap or both, although it is impossible to know how many. One goal of the study was to capture insights from rank and file AEA members that might confirm, complement, extend, or even oppose those expressed in the Roadmap. For this reason, 27 members of the AEA Board, Political Action Committee and Evaluation Policy Task Force were excluded from the initial invitation to brainstorm evaluation policy ideas. (Note that, in May of 2010, a year after the data for the present study were collected, AEA opened to all members a one-month comment period on a slightly edited version of the Roadmap, and in October of 2010, the document received formal approval from the membership with little substantive change).

Since 2009, several U.S. federal government initiatives have sought to improve evaluation, including the creation of the position of Chief Performance Officer with government-wide oversight of performance measurement and improvement efforts, the replacement of the PART initiative, whose emphasis was the grading of programs, with a Performance Improvement and Analysis Framework (PIAF), whose emphasis is agency based and government-wide goal-setting. Federal budgets have included special funding to encourage impact evaluation, and more recently, evaluation capacity building in federal agencies. There is a coordinating committee for performance measurement and a separate interagency program evaluation working group whose charge includes the development of government-wide guidance on evaluation (Congressional Research Service 2011). So far this guidance has yet to surface. AEA's "Roadmap" has since been cited in congressional testimony, in Government Accountability Office (GAO) reports, and in other U.S. Federal government settings (American Evaluation Association 2011).

Yet despite ongoing efforts to improve evaluation and its link to decision-making, the U.S. federal government still lacks a comprehensive evaluation policy, such as Canada's government-wide "Policy on Evaluation" (Segsworth 2005), or the European Union's "Financial Regulations", which specify the scope, purpose, timing, and use of evaluations (Stern 2009) or the OECD Development Assistance Committee "Principles for Evaluation of Development Assistance" (OECD 1992). While the U.S. has a performance measurement framework in the Government Performance Review Act of 1993, its implementation has focused almost solely on performance measurement, and has yet not been fully implemented government-wide for either performance measurement or evaluation (United States General Accounting Office, 2010). At the agency level, there are few freestanding, explicit policies on evaluation, although one notable exception is the recently announced Department of State Evaluation Policy (United States Department of State, 2011). However, for the most part, to find indications of an agency's intentions for evaluation, one must comb through annual agency budget requests to Congress. Across the U.S. federal government, evaluation is conducted and used in a number of different ways, and responsibility for evaluation is situated at various different levels within different branches of government and federal agencies. The work level and influence of each evaluation unit depends largely on the particular individuals holding key leadership positions at a given point in time (Grob, 2010). Major program initiatives carry their own evaluation requirements, such as No Child Left Behind, whose benchmarks for students and assessment requirements for teachers have become the subject of intense debate.

In a January 2011 report, the Congressional Research Service traces the evolution of the Obama administration's agenda for improving government performance since 2009, and offers a set of critical questions Congress could pursue in order to strengthen the government's evaluation system. Among CRS' criticisms of the Obama performance agenda and its implementation since 2009 are: 1) insufficient attention to evaluation system-building concerns, such as evaluation capacity in the government workforce, 2) an overemphasis on performance measurement at the expense of true evaluation and 3) a failure to fully articulate how the different frameworks and decision-making bodies associated with performance measurement and evaluation are to relate to one another as parts of a single, coherent whole.

Statement of the problem

The lack of a coherent, comprehensive evaluation policy framework leaves the U.S. federal government vulnerable to variations and gaps in the scope, quality and transparency of its evaluation activities over time, across different levels in the organizational hierarchy, and across federal agencies and legislative initiatives. Where there is a lack of transparency about what the rules are for evaluation and who makes them, there is the potential for exclusion and the dominance of privileged interests. Where there is a lack of coordination in evaluations across agency evaluations, results cannot be meaningfully compared. Where there is a lack of consistency in evaluations over time, trends cannot be identified showing how programs and policies are changing in their effectiveness and impact. The absence of clear evaluation policies, not just at the agency level but government-wide, makes it difficult for affected constituents to see how evaluation is (or is not) being conducted and how results are (or are not) being reported and used, in order to point out problems or demand improvements in the system. The potential of evaluation to coordinate and streamline the myriad of overlapping federal programs remains only partially realized.

The last systematic assessment of the evaluation function across the U.S. government took place over a decade ago (GAO 1998). For a clear understanding of where the gaps, overlaps and conflicts in its evaluation policies might be, the U.S. federal government needs a fresh and comprehensive assessment of its evaluation policies (both explicit and implicit) based in a well-grounded theoretical framework articulating what evaluation policy is, and what types of evaluation policy would be needed for a complete and well-functioning evaluation system. Previous theoretical taxonomies lack a connection to specific instances of the phenomenon, which would help to refine and operationalize the construct for applications in inventories of evaluation policy and evaluation policy theory building.

Purpose of the study and research questions

The main purpose of this study is to generate a grounded taxonomy of the construct of evaluation policy, including categorization criteria and illustrative examples. To that end, this study will address the following questions: 1) In the U.S. federal government context, what is considered "evaluation policy"? 2) What are all the relevant types of evaluation policy in this context? 3) How do the results of this study compare with previous framings of the construct of evaluation policy and its types, specifically the intuitive taxonomy of Trochim (2009) and the AEA "Roadmap for a More Effective Government"?

Potential applications of the taxonomy are as the basis for evaluation policy inventories and the construction and testing of theories of evaluation policy influence.

Sample, data and methods.

The population chosen for this study was the membership of the American Evaluation Association (AEA). AEA was chosen because it is the largest professional association dedicated to evaluation in the United States, with over 6,000 members representing all 50 states and over 60 foreign countries (American Evaluation Association 2011). Formed in 1986, the AEA publishes the *American Journal of Evaluation* and *New Directions for Evaluation*. When asked for non-identifying descriptors at the start of the survey, 36% of all study participants stated their main work activity as "evaluation", 15% said "management or administration", 11% said "consulting" and 11% said "research". Thirty-six percent reported their primary work setting was a college or university, 17% said a "private business" and 17% said a "non-profit organization".

Eighty percent reported having a primary residence in the United States, while 11% reported a primary residence outside the U.S..

The initial data for this study was a set of 920 open-ended survey responses in the form of evaluation policy suggestions. These come from 554 participants who responded to a May 2009 invitation to 2,000 randomly sampled members of the American Evaluation Association. The next set of data consists of two types. One is a set of statement rating inputs by participants who responded to an October 2009 invitation to 400 randomly selected AEA members. Sixty-three provided ratings of evaluation policy statements on a "merit" scale and 48 provided ratings of evaluation policy statements on a "feasibility" scale. The remaining data was a set of sorting inputs from participants who responded to an invitation to 400 randomly sampled members and 27 purposefully selected leaders of the American Evaluation Association. Twenty sets of sorting inputs are used in this study.

The initial 920 evaluation policy ideas were first categorized using a constant comparison coding approach, examined and then winnowed down to a set of 100 for sorting and rating. To analyze the data, this study uses the concept mapping method of Trochim (2009), which applies multidimensional scaling and hierarchical cluster analysis to participant sorting inputs to generate a concept map in x,y space organized by conceptual clusters. A stress value is computed for the model and its reliability is tested. Participant rating inputs are combined with sorting inputs and patterns are examined. Results are compared with categories in the Trochim (2009) evaluation policy taxonomy and the American Evaluation Association "Roadmap for a More Effective Government" (2009). An enhanced version of the Trochim taxonomy is proposed.

Results are synthesized with literature to construct a theoretical taxonomy of evaluation policy based on the dimensions of mechanisms and values, and a suggested inventory instrument is offered.

Definition of terms

Evaluation. Evaluation, in its more general, everyday meaning, of determining or fixing the value of something, has been in existence for many centuries. However, evaluation as a professional occupation, based in social science disciplines, is a relatively new phenomenon. In the United States, the profession first emerged as distinct from its parent disciplines of education, psychology, philosophy and sociology (Yarbrough, Shulha and Caruthers 2004) in the 1960s, as a response to the proliferation of Great Society programs. (Carman, Fredericks and Introcasco 2008). In its first incarnation, evaluation was a blending of economic public management and applied social science methods, such as survey research and large scale statistical analysis. United States government funding for evaluation surged in the 70s, then saw drastic cutbacks in the 80s and 90s, in tandem with weakening support for government programs. However, with increasing statutory and regulatory mandates beginning with the Government Performance Regulatory Act of 1993, the demand for evaluation continues to grow, in some cases outstripping the now limited capacity of federal agencies (Rist and Paliokas 2002). Perhaps in recognition of this fact, President Obama announced and began implementing, in his first year as president, plans to improve and enhance the evaluation of federal policies and programs. (OMB 2009)

Evaluation employs a variety of social science methods (qualitative, quantitative, mixed) to pursue different kinds of information (e.g. needs assessment, modeling, implementation and outcome information) in a range of different contexts (academic, government, non-profit, private enterprise) and from a range of different perspectives (internal, external, both). Evaluation can also be considered a "transdiscipline", providing tools to other disciplines, but also enjoying a stand-alone status (Scriven 1998). There is within the field a rich diversity of (sometimes competing) prescriptive theories of practice (Shadish, 1998). This diversity of methods, roles and theories has at times contributed to difficulty in defining evaluation as a unified profession (Stevahn, King, Ghere and Minnema 2005).

This difficulty is perhaps reflected in the many different definitions of evaluation from within the field. However, a review of these by Geva-May and Pal (1999) reveals several common features, namely: *systematic methods* (Nagel and Freeman, 1975; Rossi, Lipsey and Freeman, 1989); *valuation or judgment* of the merit or worth of an object (Joint Committee 1981, Eisner 1979, House 1980, Scriven 1966) and a *comparative* aspect, as in assessments of two or more different approaches to a problem (Alkin and Ellett, 1990). Other common features include an aspect of *feedback for improvement*, including both formative functions (Scriven, 1966) and monitoring functions (Chelimsky, 1985) and an implied *role in decision making* (Cronbach 1963, Stufflebeam et al.,1971).

Definitions of evaluation vary in the breadth and narrowness of purpose they assign to evaluation. Some argue for an inclusive view, incorporating monitoring and performance measurement under a larger umbrella of "evaluation" activities (see for example *New Directions for Evaluation*, 1996, Vol. 71, entire issue). Trochim (2010) departs altogether from the judging aspects so central to many definitions of evaluation, suggesting evaluation includes any activity

involving "… the systematic acquisition and assessment of information to provide useful feedback about some object". The definition of program evaluation inside the U.S. Government Performance Review Act (1993) narrowly focuses on performance measurement. In GPRA, "'program evaluation' means an assessment, through objective measurement and systematic analysis, of the manner and extent to which Federal programs achieve intended objectives".
On the other hand, others assert a rather specific set of functions for "evaluation", wholly distinct from and complementary to those of performance measurement (GAO 2005). According to the GAO, while performance measurement monitors the achievement (or non-achievement) of pre-specified goals, it cannot explain the reasons for these outcomes. Evaluation's distinct role, according to this definition, is to address questions of whether outcomes are, in fact, attributable to programs or policies at all, through what mechanisms programs or policies did (or did not) achieve their goals, and whether different program or policy configurations would work better.

One notable exception to this lack of unanimity within the field of evaluation is a set of well-known standards for educational evaluation. In 1975, the Committee on Standards for Testing and Use in education, composed of representatives from the American Educational Research Association, the American Psychological Association and the National Council on Measurement in Education formed a subcommittee on evaluation called the Joint Committee for Standards on Educational Evaluation. The JCSEE sought input from twelve different professional associations working in evaluation or in fields related to evaluation, and first published the standards, under the four general headings of "utility", "feasibility", "propriety" and "accuracy", in 1981 (Yarbrough, Shulha, Caruthers 2004). The standards were revised in 1994 and again in 2011. Despite the broadly inclusive process through which these standards were developed and revised, some argue that such standards are inherently value-based, and should not be universally applied. Controversies remain in the international context as to whether there exist any truly “universal” standards for evaluation, or whether more tailored “open” standards, reflecting the local values and culture where they are to be applied, are more defensible (Russon 2004). This study takes the position that it is possible to develop general evaluation standards that can be usefully applied across a large and complex organization with divergent sub-cultures if the standards are sufficiently general and flexible. However, such standards should be developed in a highly participatory process in which relevant stakeholder groups are represented, and should not be applied beyond the general context for which they were developed.

Another contended aspect of evaluation as a professional activity is the extent to which it properly includes offering specific recommendations to policy makers, or promoting the use of evaluation results. While some argue for a more neutral, “accountant” role of simply providing objective information, leaving the comparative analysis of policy alternatives to policy analysts, (Geva-May and Pal, 1999), others suggest that evaluation as a field can and should take a more proactive role in assuring that the results are used to inform program and policy decisions, and go so far as to suggest strategies by which evaluators can work to increase evaluation use (Chelimsky 1986, Grob 2003, Henry and Mark 2003). Without taking a position on the appropriateness of advocacy as an activity of evaluators, this study assumes that evaluation practices, as well as practices regarding the use of evaluation results, are both appropriate objects of public policy, and that evaluators constitute a major stakeholder group whose views and expertise should be considered in formulating such policy.

This study assumes that study participants accept or at least are aware of the definition of evaluation embraced by the American Evaluation Association, since all of the participants in the present study were solicited from among the membership list of AEA.
The following is therefore the definition most likely to be familiar to study participants:

"Evaluation involves assessing the strengths and weaknesses of programs, policies, personnel, products, and organizations to improve their effectiveness." (AEA 2010)

The AEA distinguishes evaluation as an assessment activity distinct from mere monitoring functions, and incorporates the judging aspects central to so many definitions of evaluation. It regards as the object of evaluation not just programs and policies, but also other objects, reflecting the diversity of its large membership. Finally, the AEA definition asserts that the primary purpose of evaluation is improvement, implying evaluation results should and do inform changes in and decisions about their objects. It specifies no particular method or approach a study must use to be rigorous enough to be considered evaluation.

Evaluation policy. The literature on program and policy evaluation contains many definitions of evaluation, and evaluation systems, but almost none of evaluation policy. In fact, very few of the works found in the literature review for the present study, even those whose object is evaluation policy, included a definition of the term. Government documents in the U.S. context often refer to the many sub-types within the larger conceptual domain of evaluation, for example, "assessment", "performance measurement" and "comparative research", to name only a few. Since one of the main purposes of this study is to empirically ground and refine the sensitizing concept of evaluation policy, it will be useful to start from a very broad and inclusive definition that leaves room to move in many different directions.

The common, dictionary meaning of “policy” is: a) “a definite course or method of action selected from among alternatives and in light of given conditions to guide and determine present and future decision or b) a high-level overall plan embracing the general goals and acceptable procedures” (Merriam Webster Online). Policy then, in contrast to decisions made in and for the present moment, involves some sort of proactive, conscious, deliberate process that considers both needs and alternatives in order to arrive at a decision. Policy involves the exercise of power in order to influence action, rather than being the mere expression of an idea. The action is to be repeated and in the future, rather than taking place only in a single, isolated instance.

What does this standard definition leave undefined? There is no minimum requirement for how far into the future the intended effect of a decision must extend in order to be considered a policy. In the U.S. governmental system, and in most organizations, policy is at least sometimes revisited and changed (sometimes almost immediately), yet it can still be considered policy, by virtue of the fact that it is intended to cover more than one instance of something.

Neither does the standard dictionary definition above specify how explicit a plan must be in order to be considered “policy”. It is easy to think of examples of policy at many levels of formality and informality, of rigidity and flexibility, fixedness and dynamism, ranging from those which are highly formal, explicit and static--perhaps the U.S. Constitution is an example of this--to those which are much more informal, implicit, and changing. Here, an example might be the approach guiding a private enterprise's product development strategy in a highly competitive and rapidly evolving technology market.

So the simplest, most inclusive definition of evaluation policy might be that it is any policy whose object is evaluation. However, since the first step of this study is to brainstorm toward the broadest possible set of examples of evaluation policy, an even more inclusive definition of evaluation policy is needed. The definition below is more flexible than the standard, dictionary definition of policy, and is likely to be familiar to study participants from previous exposures at recent professional conferences and in journal articles.

"Evaluation policy is any rule or principle that a group or organization uses to guide its decisions and actions when doing evaluation." (Trochim 2009)

This definition leaves room for varying levels of application and specificity, from the highest level, general principles to the lowest level, specific, step-by-step directions, including everything from high-level Congressional statements of intent to laws and regulations to the most specific procedural guidelines. This definition also leaves room for policies whose direct object might not be evaluation, but which nonetheless influence evaluation-related decisions and actions. According to this definition, evaluation policies are inherently collective--an individual establishing intentions for private practice cannot be said to be setting "policy". Evaluation policies may be written and explicit, or unwritten and implicit. They may be--and in fact, often are--adopted without having considered alternatives. The definition above widens the scope of the construct to include those unspoken or implicit rules-of-thumb which may have the effect of guiding action, but which may not have been consciously selected through any sort of deliberate process.

There are, however, boundaries on this definition. Trochim distinguishes evaluation policy from evaluation theories, standards or guidelines by asserting that evaluation policy is actually applied to guide actions or decisions in a way that includes consequences for non-compliance.

His definition implies other limitations, with the potential to constrain participant responses in this study in unintended ways. By referring to “rules”, this definition evokes the “stick” of Vedung’s “carrots, sticks and sermons” policy instrument typology (Vedung 1998). In this scheme, “carrots” are policies with incentives, sticks are policies with negative consequences imposed to enforce them, and sermons are explanations or urgings designed to motivate a particular audience, such as a speech by a high-ranking official. To the policy instrument classification scheme of Vedung, Stern (2009) adds three more policy instrument types which emerged in the European Union since the 1990s: 1) *Co-regulation initiatives* from the EU central commission that specify broad intentions, but leave detailed regulation to national authorities; 2) *self-regulation*, for example, by trade associations and through voluntary agreements and 3) more developed forms of *consultation and dialogue* intended to improve the participation of interested parties and to make consultation more transparent. Datta (2009) adds that policy is also expressed or implied in *appointments* to cabinet and other leadership posts in government. Trochim’s emphasis on “principles” and “rules” could lead the reader to believe that other types of policy instruments are excluded from the domain of evaluation policy. Last, by stating that the object of evaluation policy is “…decisions and actions <u>when doing evaluation</u>”, Trochim’s definition (perhaps unintentionally) limits the scope of evaluation policy to matters directly involved in the planning, implementation and reporting of evaluations. A literal reading of this definition would exclude activities that could not be considered “doing evaluation”, such as establishing organizational structures for evaluation or building evaluation capacity through training and technical assistance. To the extent that study participants take this definition as a foundation and then push beyond it, a more elaborated, refined and inclusive construct will be the fruit.

The next chapter situates this study in the context of the history of evaluation in the United States and in previous work and thought on evaluation systems, and summarizes recent calls for an empirically grounded taxonomy of evaluation policy.

CHAPTER TWO: LITERATURE REVIEW

The following literature review provides: 1) a brief history of the field of evaluation and U.S. government evaluation policy; 2) a summary of arguments for the conscious design of an evaluation system with key components and indicators of success; 3) recent theoretical work to define the concept of evaluation policy and its sub-types, with calls for new work to test and ground these theoretical beginnings.

Evaluation in the United States

A brief overview of the history of the field of evaluation, evaluation policy, and a description of relevant institutional structures in the U.S. federal government explain why questions about what constitutes credible evaluation evidence are the object of evaluation policy debates today, and why a more coordinated, whole-system approach to developing evaluation policy is needed.

The field of evaluation emerged in the United States in response to the exceptional proliferation of social programs that accompanied post World War II rebuilding and the Great Society programs of the 1960s (Rossi, Lipsey, and Freeman, 2004; Sanders, Fitzpatrick and Worthen, 2004; Shadish, Cook, and Leviton, 1991) Facing public demand for accountability for expenditures, the U.S. government sought new ways to measure program effects, and allocated unprecedented amounts of funding for assessments of programs. The field of evaluation did not yet exist. Researchers from social science from education, psychology, philosophy, and sociology attempted to adapt their research methodologies to the field. Evaluators also faced pressures from the Executive, Congress, and many other parties with a stake in the results of evaluations, and resistance to the use of evaluation results in decision-making (Yarbrough 2004).

In the 1980s, President Reagan presided over deep cuts to social programs and to evaluation staffing and activities in the U.S. federal government (Rist and Paliokas 2002). Federal agencies began conducting only those evaluations they could accomplish in-house. The focus of evaluation work shifted to documenting the internal management and operations of programs, away from broader policy questions of program utility and impact. Congress reacted by turning more to the General Accounting Office for independent evaluative information. In 1980 the GAO created the Institute for Program Evaluation, later renamed the Program Evaluation and Methodology Division (PEMD). However this large and well-respected evaluation unit had dwindled significantly in size and scope by the early 90s (Grasso, 1996). It was against this backdrop, that in 1986, three separate U.S. evaluation professional groups, the Evaluation Network, the Maryland Evaluation Association and the Evaluation Research Society, merged to become the American Evaluation Association (AEA) "to improve the theory and practice of evaluation". (Yarbrough 2004)

Enter President Clinton and the New Public Management movement, bringing a renewed focus on results-oriented accountability, with the Reinventing Government initiative and the enactment of the Government Performance and Results Act of 1993. These developments shifted the focus of reporting from the mere documentation of activities back to the measurement of performance against stated objectives (Rist and Paliokas, 2002), but neglected fuller evaluation studies asking deeper questions about the causal links between programs and observed outcomes, which had come to be viewed as "an expensive luxury" (Scheirer and Newcomer, 2001).

The Government Performance and Results Act (GPRA) of 1993 attempted to increase the flow of performance information and evaluation results to Congress and the President, holding federal agencies more accountable for their management of programs. Under GPRA, each agency head is required to submit to the U.S. Office of Management and Budget (OMB) a five-year strategic plan including performance goals and objectives in quantifiable, measurable form, and report on external factors that might interfere with effectiveness and progress toward goals. Despite the new statutory requirements on agencies to increase and standardize performance measurement, by the late 90s, staffing capacity for such activities still remained low, and partly as a result of this, implementation remained incomplete (GAO 2005, GAO 2010).

To complement GPRA by generating more performance information at the program level, President Bush introduced the Program Assessment Rating Tool. The PART, promoted as a consistent approach to rating federal programs, contained a standard set of questions about whether a program's design and purpose were clear and defensible, whether agencies were setting valid annual and long-term goals for programs, whether programs were well-managed, and whether program results were reported with accuracy and consistency. The PART provided for a scoring system that yielded ratings of "Effective", to "Moderately Effective", to "Adequate", to "Ineffective", or "Results Not Demonstrated".

Because they were not conceived as part of a single, coherent system built with input from all branches of government, GPRA and PART struggled to achieve their intended objectives, and became battlegrounds for power struggles between the Executive and Legislative branches of government. Training large numbers of federal agency staff from different agency contexts to consistently implement the PART across diverse programs proved extremely difficult. Congress, had not been consulted in the construction or implementation of PART, and expressed concerns that the President was injecting ideological bias into evaluation processes so as to further his own agenda (GAO, 2004). Federal agencies struggled to comply with overlapping and conflicting reporting foci and measurement requirements of GPRA and PART (GAO, 2005b).

By this time, the struggles of evaluators to reconcile the agendas of stakeholders with the imperative to generate high quality, rigorous evaluations had lead to a division within the field of evaluation around the role of evaluators in designing evaluations and selecting methods. According to Datta (2011), in the so-called "populist" view, that sees "…evaluators as taking into account these multiple voices, including (for some evaluators) being advocates for the most disenfranchised" (p.279). For these evaluators, evaluation is a political act, and collaborative, participatory and so-called "social justice approaches" are the standard. The evaluator facilitates negotiations among stakeholders in the development of the evaluation design. On the other hand, the so-called "public interest" view holds to "…a belief in the public interest or common good that transcends the diversity, a highest common denominator, and a role for evaluators as sources of unbiased, fair information relating to this interest." In general evaluators in this camp favor rigorous evaluations according to pre-set standards (ibid, p. 281).

Shortly after PART's debut, the Bush administration endorsed the use of randomized controlled trials as the gold standard for PART-related evaluation studies. Representatives from a group called the Coalition for Evidence-Based Policy initiative strongly supported the move, arguing that "evidence of effectiveness generally cannot be considered definitive without ultimate confirmation in well-conducted randomized controlled trials" (Coalition for Evidence-Based Policy, 2009). However, the move to privilege RCTs in PART assessments sparked staunch opposition from evaluators from the American Evaluation Association and the American Education Research Association, who argued instead for acceptance and recognition of a continuum of different evaluation methods, from which an approach should be chosen depending on the needs of stakeholders and the stage of development of the program or policy under study (Chelimsky 2007, Julnes and Rog 2007, GAO 2009). The American Education Research Association (AERA) would later develop and promote a formal definition of "scientifically-based research" for use by Congress in writing legislation. The AERA definition explicitly endorses, for examining causal questions, a variety of methods, including "random assignment or quasi-experimental or other designs that substantially reduce plausible competing explanations for the obtained results, such as longitudinal designs, case control methods, statistical matching, or time series analyses" (American Educational Research Association, 2010). The American Evaluation Association would later develop a set of principles and guidelines for federal agencies to use in developing evaluation policy, among which is the independence of evaluators in selecting evaluation designs and methods (American Evaluation Association "Roadmap for a More Effective Government", 2009)

Due to the many implementation challenges and controversies surrounding PART, OMB stopped conducting program-specific PART reviews in 2009. Despite the ambitious goals of GPRA, implementation across large and diverse federal agencies has proven difficult in the face of the problems of low evaluation and the lack of collaboration between the Executive and Congress.

PART’s focus on rating the performance of individual programs and GPRA’s focus on goal-setting at the agency-level planning left a vacuum in the area of higher level, cross-agency analysis of programs and policies (GAO, 2004). While GPRA’s framework includes a requirement for a “government-wide performance plan”, this provision, according to the GAO, its potential for helping address government-wide problems with mission fragmentation and program overlap has never been fully realized (GAO, 2000b; GAO, 2010).

Difficulties in developing a government-wide framework for evaluation, including both performance measurement and fuller evaluation studies, are partly a function of the checks and balances built into the structure of the United States federal government. The Executive branch’s program mangers conduct ongoing monitoring and analysis of programs, while Congress conducts less frequent reviews for broad oversight of government programs and policies. The Office of Management and Budget is responsible for overseeing evaluation in the Executive. Each Executive agency also has an Inspector General’s office, charged with “audit and oversight”, whose scope is limited to preventing fraud, waste, and mismanagement in programs with federal funding. Many agencies also have a freestanding evaluation unit. For Congress, four legislative support agencies provide evaluative information—the Government Accounting Office, which takes on studies directed by Congress and can also initiate studies, the Library of Congress and Congressional Research Service, which respond confidentially to requests for information about national issues, and the Congressional Budget Office, which provides information to Congress on the budget and generates fiscal impact estimates for legislation. Congress and the President also receive evaluative information from many other sources, including lobbyists, special interest groups and think tanks (Rist and Paliokas 2002).

While intended to prevent any one branch of government from becoming too powerful, and to promote critical thinking, according to the U.S. G.A.O., the decentralized structure of the U.S. government has also lead to agency-specific, programmatic silos, and problems of multiple overlapping yet uncoordinated programs in different agencies. (GAO 2010)

In the face of these challenges, evaluators have continued to argue for a renewal of evaluation in the U.S. federal government, asserting that evaluation is not a "luxury", but rather an essential component of an effective performance management system. Wholey (2001) argues that to achieve the goals of results oriented management, mere performance measurement is not enough. Rather, it is first necessary to develop a consensus among key stakeholders on the mission, goals and strategies of a program or policy, then to move to measuring outcomes, and then proceed to using performance information to improve programs. According to Wholey, the role of performance measurement is to motivate managers to use evaluation information to improve performance and communicate the value of agency activities to stakeholders and the public. The complementary role of evaluation is to give meaning and force to performance data by designing high quality measures and measurement systems, and by conducting studies addressing not just whether programs and policies hit their targets, but how and why (or why not). Such studies would measure program implementation, test causal assumptions, explain the reasons behind current performance levels, measure unintended outcomes, assess the cost effectiveness of programs, estimate causal impact or net benefits, measure hard-to-measure outcomes, and make recommendations on how to improve performance (p. 345)

Moving in this direction, recent Presidential initiatives have attempted to make performance information generated under GPRA more useful, increase the use of deeper causal analyses from impact evaluation studies, support a somewhat broader range of evaluation methodologies, and foster cross-agency coordination. The 2010 Presidential budget issued a call to each federal agency to identify a set of high priority performance goals as the first step toward establishment of a new "performance improvement and analysis framework" to replace PART. Agencies are now being required to submit an annual performance plan explaining strategies to move toward their stated goals, including any shifts toward more evidence-based approaches, and to report not just performance snapshots, but trends in performance including a detailed explanatory analysis. (OMB, 2010) The President's 2010 "Evaluation Initiative" allocated $100 million in grants for new evaluation studies as part of a strategy to motivate agencies to conduct more high quality evaluation studies and to use evaluation results. In subsequent budget proposals, this initiative has been renewed and expanded to include funding for evaluation capacity building. The President also charged an interagency working group with the broad mission of building agency evaluation capacity, fostering evaluation networks, and developing government-wide guidance for evaluation practices (United States Presidential Budgets, 2010, 2011 and 2012).

In a January 2011 report, the Congressional Research Service traces the evolution of the Obama administration's agenda for improving government performance since 2009,and offers a set of critical questions Congress could pursue in order to strengthen the government's evaluation system. Among CRS' criticisms of the Obama performance agenda and its implementation since 2009 are: 1) insufficient attention to evaluation system-building concerns, such as evaluation capacity in the government workforce and 2) a failure to fully articulate how the different frameworks and decision-making bodies associated with performance measurement and evaluation are to relate to one another as parts of a single, coherent whole.

Evaluation systems

The following is a summary of arguments for the conscious design of an evaluation system, and a set of suggestions from the literature for what to consider in designing one.

According to Dahler-Larsen (2006), evaluation systems serve a crucial function in providing "continuous streams of information that play an integrated role in the functioning of an organization or a field of organizations" (p.1). In this view of the evaluation system as a type of information system, essential elements of a successful evaluation information system include: 1) an organizational structural unit for evaluation; 2) evaluation criteria; 3) information technology to collect and retain data; 4) a self-representation that justifies the system; 5) social support, including financial, political, sufficient knowledge, skill and motivation among implementers of the system and among those charged with collecting data and 6) alignment of all these factors. Evaluation policy would address all of these elements in a complete and integrated evaluation system.

The Organization for Economic Cooperation and Development Development Assistance Committee (OECD DAC) affirms the power of evaluation systems to manage the flow of evaluation results to and promote their use by decision makers. DAC recognizes that the full use of evaluation evidence for decision making depends upon "…the evaluation function's independence, the resources it relies upon and, not least, cultural attitudes to evaluation." (OCED DAC, 2004, p.4) Speaking also from the development aid context, Liverani and Lundgren (2007) take a larger view of the evaluation system as including "…the procedural, institutional and policy arrangements shaping the evaluation function and its relationship to its internal and external environment…" (p. 241). In their view, designers of evaluation systems should consider forces both within and around the organization, including the supply of and demand for evaluation.

Leuuw and Furubo (2008) acknowledge the rise of the evaluation system as part of a growing trend away from a focus on individual evaluations and toward "systematic reviews, syntheses, information systems, systems of good practices, m(onitoring) and e(valuation) systems, performance monitoring, inspection and oversight, repositories of evaluation results and observatories" (p. 158). From a synthesis of prior work, they assert a set of simple criteria that define this phenomenon. According to Leuuw and Furubu, an evaluation system must have a distinctive epistemological perspective, representing an internal consensus about what evaluation is, why the organization should do it and what kind of knowledge to produce. Second, responsibility for evaluation must rest within the organization, rather than being contracted out. Third, there must be a tendency toward permanence--replacing ad hoc evaluation activities with planned ones. Fourth, results must be linked in some way to decision making about and implementation of policies and programs. These authors argue that evaluation systems embedded within larger administrative systems such as governments tend toward the production of "routinized information for day-to-day administrative practices", but "little [information] of relevance for fundamental reassessment" (p. 165). They also argue that those working within administratively embedded evaluation systems tend to perpetuate these systems for the sake of their own careers and reputations. He argues for careful and conscious design of evaluation systems so as to avoid these perverse effects.

In a similar vein, Segerholm (2003) focuses on the internal aspects of government contexts and their effects on evaluation systems. She emphasizes the influence on evaluation practice of ways in which the programmatic work of an agency is structured, staff's views on and knowledge of evaluation, and internal power relations. She also points out that evaluation systems and political contexts tend to interact in ways that are mutually reinforcing, and calls for critical re-assessments of existing evaluation systems.

Schaumburg-Muller (2005) offer a way to critically view evaluation systems as playing different roles in different types of organizations. In their typology, the *learning organization* type is a modern participative and innovative multidisciplinary organization that has a professional culture of commitment and results and uses evaluation to reap learning for innovation and development. It assumes a homogenous and flat organization with evaluation integrated into operations, and also assumes that evaluations are both of high quality and relevant. In the *functional/rational* type of administrative system, actors behave rationally to achieve goals and use evaluation results to make decisions. In doing so, it assumes high quality information, clear goals and clear causal relations in results. Evaluation is used for documentation and accountability to legitimize the organization to outside stakeholders. The *political organization* type operates from a loose network of power relationships, with each actor or group having their own agenda, ideologies and goals. The organization makes decisions based on power, coalitions and compromises.

Evaluation is not used for learning, but rather as a political instrument in power struggles. Independent evaluation is mistrusted. The *institutional* type is driven by norms, values, procedures (not a higher set of objectives), and consists of cognitive, normative and regulative structures and activities carried out in cultures, structures and routines. It is more concerned with procedures and appearances than instrumental utilization of evaluation. The focus is on evaluation means and procedures rather than on use of results. In this type of organization, leadership shows its commitment to be accountable by specifying guidelines, following international standards and by demonstrating that it has norms and rules for its evaluation work (p. 16). While it's unlikely an organization would consciously adopt policies toward becoming an institutional type, the description here may present a useful mirror for organizations taking inventory of their evaluation system.

Chelimsky (2009) describes the ways in which professional culture and bureaucratic climate of a government agency and larger, surrounding governmental structures and ideologies can constrain evaluation practices. She argues for more intentional design of evaluation policies to directly and systematically address the problems generated by this context. She acknowledges that many of the contextual or external components of evaluation systems, such as the historical relationship of information to decision-making in the organization, change slowly and are less easily manipulated using policy levers.

The table below summarizes key components of evaluation systems from this review.

Table 1: Components of evaluation systems

	Internal	**External**
Dahler-Larsen (2006)	distinct evaluation unit evaluative criteria IT resources financial, political support staff knowledge & skill staff motivation	financial support political support
Liverani and Lundgren (2007)	procedures, institutions, policies relationships with internal actors supply of (internal) evaluation	organizational culture relationships with external actors supply of (external) evaluation demand for evaluation
Leuuw and Furubu (2008)	epistemological perspective internal capacity organizational permanence	links to decision making (use)
Segerholm (2003)	structure of programmatic work staff views and knowledge internal power relations	
Shuamberg-Muller (2005)		purpose of evaluation within the organization
Chelimsky (2009)	professional culture bureaucratic climate governmental structures structure of program work staff knowledge and attitudes	

A review of meta-studies of evaluation systems suggests that the institutionalization of evaluation functions within government and a coordinated approach across the evaluation system are considered important factors predicting effectiveness in government performance measurement and evaluation.

The Development Assistance Committee (DAC) of the Organization for Economic Cooperation and Development (OECD) was formed as a forum for donor countries, and has grown to assume an important role in coordinating donor country evaluation policy (Debelstein and Rebien 2002). By 1991, the Development Assistance Committee published its “Principles for Evaluation of Development Assistance” to guide donor agencies in establishing strong, central evaluation policies (OECD DAC, 1992). These principles call on donor agencies to set forth clear roles and responsibilities for evaluation, assign evaluation to a strong position within the agency, require clear articulation of purpose of the evaluation within the agency, and provide for the integration of evaluation with planning. In subsequent reviews of DAC member agency compliance with the principles, DAC’s criteria for success emphasized the degree of the evaluation function’s integration within governmental systems of the agency, its stability over time, and the extent of coordination among units charged with evaluation (OECD DAC 1998; Liverani and Lundgren 2007, Foresti, Archer, O’Neil and Longhurst, 2007)

Mackay (2007) of the World Bank articulates criteria for successful “monitoring and evaluation (M&E)” performance measurement systems which include the extent to which the M&E system is institutionalized, in the sense that it is not dependent on the sponsorship of particular government officials or private donors, and the degree of harmonization and coordination within the system, avoiding duplication of effort.

The European Union (E.U.) Financial Regulations regulated the scope, purpose, timing and use of evaluation in member countries (Stern,2009). They set a minimum level of evaluative activity for member countries, and require that all Directorates General have their own evaluation functions or units to lead evaluation activities. Each country must present an annual plan indicating which programs will evaluated to the E.U. College of Commissioners, perform an annual inventory of all evaluation reports, and oversee a network of evaluators for conducting peer reviews (1996 Communication on Evaluation). Assessments of compliance and implementation of the Regulations among E.U. member countries has focused on the extent to which a country's evaluations meet standards for quality, the scope and direction of evaluations, the involvement of stakeholders, and the extent to which evaluation results integrated into decision making processes (Summa and Toulemonde, 2002).

Furubu, Rist and Sandahl (2002) conduct a comprehensive comparative study of "mature" country-level evaluation systems in twenty-one countries around the world. Their rating scheme highlights both institutionalization and pluralism of the evaluation function. Countries in the study include the United States, Canada, Australia, Sweden, the Netherlands, the United Kingdom, Germany, Denmark, Korea, Norway, France, Finland, Israel, Switzerland, New Zealand, Ireland, Italy, China, Spain, Zimbabwe and Japan. The authors synthesize studies of the policies, institutions and practices within each country to arrive at a rating score indicating the "maturity" of that nation's evaluation system. The authors' criteria for a "mature" evaluation system include conducting evaluation from multiple points of view, as well as the integration of evaluation into all of the country's political and administrative systems. Items on the "maturity" rating scale include: 1) the presence of evaluation in many domains in the public sector; 2) a diverse supply of competent evaluators; 3) a national discourse on evaluation that is specific to national circumstances; 4) professional evaluation associations; 5) permanent institutional arrangements in the executive and 6) permanent institutional arrangements in the legislative branches for conducting evaluation and disseminating results to policy makers; 7) a diversity of different entities commissioning evaluation studies in each domain; 8) an important role for evaluation in the supreme audit function and 9) a focus on outcome evaluation. This rating scheme values pluralism and diversity in ways of approaching evaluation over integration of evaluation in a coordinated system. In this scheme, institutionalization is considered compatible with pluralism.

The table below summarizes indicators of successful quality evaluation systems found in these meta-studies.

Table 2: Indicators of a successful evaluation system

	indicator
OECD DAC 1998; Liverani and Lundgren 2007, Foresti, Archer, O'Neil and Longhurst, 2007	integration stability horizontal coordination
Mackay (2007) World Bank (M&E systems)	institutionalization coordinated approach match of supply and demand
Stern (2009) Summa and Toulemond (2002)	quality of studies scope and direction use in decision making stakeholder involvement
Furubu, Rist and Sandahl (2002)	institutionalization of evaluation broad-based evaluation diversity and pluralism

Taxonomies of evaluation policy

The idea of a conscious, systems approach to designing a set of evaluation policies for a more effective government has deep roots and some history. So what? What works in other contexts won't necessarily work in the U.S. context. However, some evaluators in the U.S. believe a systems approach, provided it allows for the pluralism of our governmental structures, is the answer. Given the history of the field in this country, it is not surprising that evaluators are arguing for a comprehensive re-design of the rules of engagement. I am one of them. As we begin, it is useful to be aware of thinking and experience about evaluation systems from other contexts.

Reflecting on the Reagan era's dismantling of evaluation in U.S. government, Cordray and Lipsey (1986) wrote ‘‘ . . . evaluation studies have been predominantly driven by external forces . . . Although contributing to the rapid development of evaluation, these . . . also have left evaluation vulnerable to the ebbs and flows of the political process. The consequences of this dependency have been both good and bad. External pressures have been good in the sense that rapid developments in methods and perspectives emerged as practitioners and theorists attempted to meet the needs of users and clients. They have been bad for evaluation practices in the sense that external forces have pressured the field into valuing one-dimensional attributes (e.g., immediate utilization versus technical/statistical quality) as a criterion for [evaluation] success and [its] continued existence’’ (p. 31). One response to this dilemma has been for evaluators to argue, somewhat paradoxically, for more explicit and thoughtful institutionalization of evaluation within government so as to achieve independent and critical thinking by evaluators.

Cooksy, Mark and Trochim (2009) argue that explicitly articulating the rules and principles governing evaluation in the U.S. federal government context can increase the transparency and consistency of the rules of engagement between evaluation and government, freeing evaluators to some extent from political pressures, and allowing them a more critical, big-picture scope. Not only should there be evaluation policy, they argue, there should be a set of mutually complementary policies, whose domain is shaped with input from evaluation practitioners. They assert: "…to guide the study and development of evaluation policy, we need to know what dimensions or topics it encompasses" (p. 104). They suggest a theoretical framework can help "define the limits" and "fill in the contours", so as to more clearly identify opportunities for improvement of evaluation policy. In listing different areas evaluation policy might address, they seek to extend the scope of evaluation policy debate beyond issues of control over choice of evaluation methodology to a broader range of issues, including "…a number of considerations, ranging from management to method to participation" (p.106).

Datta (2009) credits recent controversies over attempts to dictate methods under PART with stimulating a useful "conversation" between evaluators, as the experts and suppliers of evaluative information, and government, as the consumers and demanders of evaluative information. Emphasizing how a taxonomy can be instrumental in furthering evaluators' advocacy agenda, she argues it can serve to focus the attention of policy makers on areas where the voice of experts is most needed. Datta cites an older, more evaluator-focused set of evaluation policy types assembled by the American Evaluation Association Evaluation Policy Task Force in 2007, with categories as follows: 1) *how evaluation is defined*, whether including the full range of ex ante and ex post types or merely performance measurement; 2) *when, what and how often to evaluate*, whether a broad, shallow approach or a selective, in-depth approach; 3) *choice of methods for evaluating different programs*, whether randomized controlled trials only or a range of methods; 4) *professional qualifications of evaluators*, whether specific evaluation training is required or not; 5) *budgeting for evaluation work*, for example whether money is set aside in project and program budgets for evaluation or whether evaluation has a separate budget line; 6) *how to implement evaluation*, when it should be done by internal personnel and when by external evaluators, and where evaluation sits in the organizational structure and 7) *ethics for evaluators*. To these, Datta adds two new areas: 1) *who is involved* in setting policy, whether there is diversity in participation by stakeholders from academic, applied, state, federal and local contexts and 2) *who gets funding*, whether there is a diversity of different types of evaluators and firms funded by government to conduct evaluations.

Trochim (2009) argues that an important practical purpose for developing an evaluation policy taxonomy is to help organizations with the practical task of assuring that their policies "address all the relevant aspects of evaluation" in a coordinated fashion (p.22). Trochim lays out an possible taxonomy of evaluation policy composed of the following eight areas: 1) *goals*, in the sense of the organization's goals or purposes in conducting evaluation, whether for accountability, learning or some other purpose; 2) *participation*, in the sense of who has a say in designing evaluation and setting evaluation policy; 3) *capacity building*, in the sense of training or technical assistance for existing staff or adding evaluation staff; 4) *management* of staff time and resources dedicated to evaluation; 5) *roles,* in the sense of who is responsible for what in evaluation activities; 6) *process and methods,* which encompasses issues evaluation design and measures; 7) *utilization* by managers and decision makers of evaluation results and 8) *meta-policies,* which covers periodic assessment of evaluation functions. These categories correspond roughly to the stages of a practical evaluation, which might be roughly summarized as: 1) decide the purpose of the evaluation and select evaluation questions; 2) decide which stakeholders to involve in doing each of these things and seek their input; 3) recruit the people who will do the work; 4) find resources for the evaluation; 5) assign evaluation tasks; 6) decide what methods best fit the program and questions, and conduct the evaluation; 7) decide how results will be reported and report them in a way that best facilitates use of the results for intended purposes; 8) assess implementation of the evaluation for quality improvement in the next cycle.

Cooksy, Mark and Trochim (2009) point out a need for additional work on the taxonomies they present, suggesting two different ways to test the frameworks they offer. They ask whether the categories are the right ones, whether they represent a complete set, how the categories might be refined, and whether they apply across various contexts. They call for an empirical check of the taxonomies against actual existing evaluation policies. Trochim argues a need to solicit the views and expertise of the broader field of evaluation as to whether his categories are the right ones, since evaluators and evaluation researchers are intimately engaged in evaluation and its study, and can therefore offer insights to ground the theory of evaluation policy in experience and research.

The next chapter outlines this study's plan for gathering survey responses from members of the American Evaluation Association in order to empirically test previous formulations of the concept of evaluation policy and its sub-types, ground them in the perceptions of a larger group of evaluation practitioners and researchers, and operationalize policy types for practical applications.

CHAPTER 3: METHODS

This chapter introduces the concept mapping method of Trochim (1989) and describes the application of concept mapping in this study.

A concept map is a diagram depicting the relationships among concepts. It was first developed as a graphical tool for organizing and representing knowledge domains (Novak and Gowin, 1984). Concept mapping approaches have their roots in constructivism, an epistemology that argues humans actively construct knowledge and meaning from their experiences. It has also been used in developing theory for use in research and evaluation, exploring understandings and knowledge within a group, organizing teaching and research and facilitating the creation of a shared vision for strategic planning (Canas, Novak and Gonzalez, 2004).

The concept mapping method of Trochim (1989), hereafter referred to simply as “concept mapping”, is a structured group conceptualization technique developed to build theoretical models directly from the everyday experiences and insights of practitioners in a given field, as “constituencies directly familiar with the phenomenon in question” (Trochim, 1985). The method was designed for conceptualization tasks in planning, such as developing organizational or program mission, goals and objectives. It has similar uses in research, such as forming a group consensus about the purposes of a study, its hypotheses, central constructs and key variables (Trochim and Linton, 1986). Concept mapping is a mixed method, employing both qualitative and quantitative data and methods, including open-ended question responses and quantitative rating inputs, content analysis, multidimensional scaling and hierarchical cluster analysis. In this method, informants generate specific examples, which they then group into piles or clusters. The clusters are named and the resulting clustered concept map is interpreted and may be used for various purposes, including building a logic model, developing an evaluation instrument, or building a strategic action plan or research agenda.

As a mixed method, concept mapping most closely resembles the sequential, interactive "development" type of Greene, Caracelli and Graham 1989. According to Greene et al, there are five distinct purposes for mixing methods: triangulation, complementarity, development, initiation and expansion. In a "development" study, the purpose of mixing qualitative and quantitative analysis is to capture the benefits of both by using the results from one to inform the development of the other. In concept mapping, results from the first process (categorization) are used as the raw material for the second process (multidimensional scaling). Here, the strengths of qualitative methods for capturing individual perceptions are combined with the strength of quantitative methods for aggregating across individuals and showing the degree to which perspectives are shared.

With informant categorizing as a central activity, concept mapping is well-suited for constructing an empirically grounded taxonomy. According to Smith (2002), a taxonomy is distinct from a conceptual typology, in that a taxonomy classifies, on the basis of empirically observable and measurable criteria, items as sharing common properties. Smith argues a central challenge of developing taxonomies is establishing consistent criteria for how to classify or categorize items, without which a taxonomy is essentially useless. If a taxonomy is to be useful for classifying relevant items outside the original data by study participants, criteria for placing items within categories in the taxonomy, known as “inclusion rules”, need to be clearly articulated. This is closely related to the challenges of developing systematic procedures for categorization in content analysis (Stemler 2001). According to Jackson and Trochim (2002), content analysis is vulnerable to bias when classification is the work of a lone researcher, since the researcher's understandings may not accurately reflect the state of knowledge as constructed by the relevant constituencies in a particular domain. In qualitative research, this problem may be addressed by various kinds of expert validity checks. Concept mapping draws from a broad base of perspectives within a relevant community of interest and systematically aggregates them, providing a kind of built-in, key informant validity check.

Related to the problem of valid and reliable classification is the problem of transparency in the development classification criteria. This problem is common not only to taxonomies, but to any research involving categorization (Constas 1992). Concept mapping participants are not asked to explain the specific criteria by which they sort statements into categories, so their individual criterion for these decisions are lost. However, concept mapping allows for a solution to the problem of transparency in categorization in two ways. First, groups or researchers can carefully document their decision process in settling on a name for each cluster. Second, the set of statements in each cluster serves to anchor the concept for the development of inclusion rules, which can be used in future applications of the classification scheme.

Overview of concept mapping method.

The concept mapping method of Trochim (1989) has four initial steps: 1) preparation, which involves engaging a group to work with and formulating a central question and rating scales, 2) generation and 3) structuring, which correspond to data collection or generation, and 4) representation, which corresponds to various analyses and their visual presentation.

In the preparation step, the researcher formulates a so-called “focus prompt”, or open-ended question for use in the generation step, to stimulate brainstorming by study participants. Also in the preparation step, the researcher generates criteria by which ideas generated in the brainstorming process will later be rated, and selects and recruits a sample of participants. Next, in the generation step, the participants respond to the focus prompt by free-listing statements that describe the subject of the prompt as they see it. In the structuring step, the participants give their sense of the conceptual “closeness” and “distance” among all of the statements by sorting them into categories, and naming the categories by theme. Also part of the structuring, participants rate each statement by assigning a number from one to four or one to five, typically on two scales: importance and feasibility. Last, in the representation step, participant inputs on conceptual closeness are aggregated quantitatively and translated into a two-dimensional dot map. Cluster analysis shows each group of statements thought by participants to express a common theme appearing as a group of dots enclosed by a geometric figure. The map may be interpreted by the researcher alone or together with study participants.

Two other analyses are often performed in concept mapping. To identify high priority items for action, the group’s ratings of statements on merit are mapped together with its ratings on feasibility in a two-dimensional graph with x and y axes. Here, statements rated highly on both importance and feasibility appear in the upper right quadrant of the graph, creating a “go-zone”. This is particularly useful for strategic planning, in assigning priority to items for immediate or longer-term action. To compare sub-group differences within the study sample, the "importance" and "feasibility" ratings provided by different demographic subgroups of participants are compared with one another in a "pattern match" graph (Kane and Trochim 2007). The following section describes the specific procedures for concept mapping used in this study.

Preparation

The American Evaluation Association is the population for this study. The American Evaluation Association is the largest professional association dedicated to evaluation in the United States, with over 5500 members (AEA Member Scan 2007). Also, unlike other professional associations, whose membership includes evaluators funded by the federal government, AEA is actively engaged in public advocacy on evaluation policy, and therefore likely to be more informed on and willing to offer ideas on this topic.

AEA publishes the *American Journal of Evaluation* and *New Directions for Evaluation* and holds an annual professional conference. According to the AEA 2007 membership survey, which had a response rate of 49%, AEA membership includes individuals with a professional interest in evaluation employed in a variety of settings, with most primarily employed in non-university settings. About half report having doctoral degrees, and most of the remaining members report masters' degrees. Education and health are the top two content areas. Professional roles include evaluation practitioner (49%), faculty member (15%) researcher (14%) and student (7%). Of those who responded to the member survey, 87% live primarily in the United States.

A principal aim of this study was identifying the broadest possible set of evaluation policies, in order to build the most complete set of evaluation policy types possible for the U.S. federal context. Accordingly, instructions to the participants in the generation or brainstorming phase of the study encouraged them to think broadly about all possible policies related to doing evaluation in the U.S. federal context.

In keeping with this aim, the form of the data generation instrument was a single, open-ended question in the form of a "focus prompt", in the form of a declaratory sentence left intentionally incomplete. Participants are to complete the sentence as many times as they wish, once for each distinct idea. They are also encouraged to look at the statements of others if they wish, so as to discourage repetition in responses and encourage the addition of novel ideas not yet provided by previous respondents. In order to avoid confusion of "evaluation policy" with the evaluation **of** policy, the focus prompt for this study is prefaced by a clarifying definition, as follows:

"For purposes of this study, an 'evaluation policy' is any rule or principle that a group or organization uses to guide its decisions and actions when doing evaluation.

The focus prompt for this study is:

"In a comprehensive set of U.S. federal evaluation policies, one policy that should be included is..."

Participants entering the study answered a short set of non-identifying, self-descriptive questions about their professional background. The questions asked about educational level, primary work setting, main professional activity and country of residence. These particular descriptors are deliberately chosen from among descriptors in the AEA membership survey, for easy comparison. The first descriptor, "highest degree", asks about participants' level of academic training, an indicator of the sophistication with which individuals understand evaluation policy. Level of education is relevant to the content of statements offered by participants in brainstorming, and the way they sort and rate statements. The second descriptor, “work setting”, reflects participants’ work-related perspective on evaluation. This is relevant as it may determine the aspects of evaluation policy with which respondents have come in contact, as well as their sense of efficacy with regard to influencing policies. The third descriptor, “major activity” reflects participants’ professional perspective, which may indicate practical experience with evaluation policy. The fourth descriptor, “country of primary residence” reflects the degree of day-to-day experience with U.S. evaluation policy. This is included because non-U.S. residents were included in the study.

Concept mapping projects typically use two rating scales: “importance” and “feasibility”. The rating instructions for this study asked participants to rate each of the 100 evaluation policy statements according to “merit”, and “feasibility”.

The “merit” rating instructions ask participants to provide a rating from 1 to 5 for each of the 100 statements to indicate whether they agreed the policy idea **should** be enacted. A bipolar Likert scale is used to allow participants to respond negatively, neutrally or positively, as follows:

> *“For each policy idea, please indicate whether you agree or disagree that it should be implemented. Rate each policy on a scale from 1 to 5 with:*
>
> *1=strongly disagree*
> *2=disagree*
> *3=neither disagree nor agree*
> *4=agree*
> *5=strongly agree.”*

The feasibility rating instructions ask participants to assign a number from 1 to 5 for each of the 100 statements to indicate whether they thought it **could** be enacted. A bipolar Likert scale is used to allow participants the opportunity to offer negative, neutral or positive rating responses, as follows:

> *"For each policy idea, please indicate whether you think it is feasible to implement. Rate each statement on a scale from 1 to 5 with:*
>
> *1=very infeasible*
> *2=infeasible*
> *3=neither infeasible nor feasible*
> *4=feasible*
> *5=very infeasible."*

Concept mapping projects typically ask participants to sort items into groups according to meaning or theme. Sorting instructions on the website directing participants to group statements as follows:

"Below are 100 policy statements (rules or principles) related to evaluation. Please place the statements in categories according to meaning or theme, in a way that makes sense to you.

Instructions:

1. First, read through the statements in the "Unsorted Statements".

2. Next, move each statement into a category. To name each category,
type a word or phrase that describes the meaning or theme. You can change where you place a statement at any time.

3. Continue to add categories and statements to categories until done. Make any final adjustments needed.

Please try not to leave any statement by itself in a category.
Please make sure every statement is assigned to a category.

Please do NOT create categories according to value or feasibility, such as "important" or "difficult". Please avoid creating categories such as "misc" or "other". "

Once questions, rating scales and instructions were complete, a request for exemption from full human subjects review was submitted to Cornell university's Institutional Review Board (IRB). The application included procedures for assuring that none of the data provided by participants could be connected with personally identifying information. These included: 1) anonymous sign-in for the brainstorming phase, and, 2) the option of designating a non-email log-in ID for the sorting and rating phases. The IRB approved an exemption from full IRB review. Though not required by the IRB, the instructions to participants include an abbreviated informed consent from each of the study participants, regardless of whether they completed the brainstorming, sorting or rating activities.

Data generation

The researcher applied to AEA for permission to use its mailing list. The formal application to AEA included a description of the study and its benefits to members. This application described the primary benefit to the organization as the opportunity for broader engagement of the membership with the topic of evaluation policy. Recruiting emails sent to AEA members included a specific disclaimer stating the data collection was part of an academic research project and not a formal, AEA Board-sanctioned process to gather member input. Recruiting emails also included an invitation to respondents to contact AEA leadership with any questions or objections to the study. The timing of emails was also coordinated to be non-overlapping with mass mailings from AEA leadership to members, and email solicitations from other researchers to members.

The AEA provided two email distribution lists, one of the entire membership (N=5,769) and one of leadership bodies including the AEA Board, Public Affairs Committee and Evaluation Policy Task Force (N=27). In May of 2009, a random sample of 2,000 members of AEA was drawn from the overall member list of 5,769 using Excel RAND function. Those in the sample received a recruiting email from the researcher inviting online participation in the brainstorming phase of the study. In October of 2009, two new random samples of 400 as-yet-un-solicited AEA members were selected. This was done to reduce the burden on individual AEA members, who receive many study solicitations from students each year, and with the thought that the study would obtain a stronger response rate from individuals who had not already given their time to participate in the brainstorming process. The first sample of 400 received an invitation email inviting online participation in the sorting phase of the study. The other sample of 400 received an invitation to participate in the rating phase. The 27 members of the Board, Public Affairs Committee and Evaluation Policy Task Force had been excluded from the initial, brainstorming sample. This was done in order to obtain ideas and sort inputs most representative of rank and file members whose views had not already been expressed in recently published AEA position papers (AEA EPTF, 2009). However, these 27 members of leadership bodies were invited to sort and rate, since they are likely to have greater interest in the topic and be generally better informed to assess the relative importance of various evaluation policies.

The sorting and rating invitations went out just before the annual AEA conference in mid November of 2009, and the opportunity to respond remained open for the entire month. The researcher made a brief presentation on the study at the conference, partly in order to stimulate interest in participation among those who had received invitations. In response, three additional AEA members attempted the sorting task in the study.

Options for the so-called “brainstorming” stage of concept mapping, in which participants respond to the survey focus prompt, include synchronous or asynchronous, in-group or online procedures. The study used a web-based approach, employing a site hosted by Concept Systems Incorporated of Ithaca, a company whose services center on supporting users of the concept mapping method of Trochim (1989). This approach was chosen to encourage broad participation by allowing participants to add data from their desktops at times convenient for them, in order to include as many participants as possible from among an organization spread out across the United States and beyond.

The brainstorming phase began with an email invitation to participate. The Cornell Survey Research Institute was hired to send emails from the researcher to each of the email addresses on the randomly selected distribution list. The advantage of this procedure is that SRI has the technical capability to send mass mailings that appear in recipients’ inboxes as individual emails, reducing the likelihood that the email will be deleted as spam.

Recipients responded according to their interest. Once participants reached the web page for the survey, they were asked to respond to the five demographic questions and then respond to the focus prompt. Participants were encouraged to submit an idea as many times as they wished. As soon as a statement was added and the participant entered “submit”, that entry was visible on a corner of the screen with other previous entries. In this way, participants were aware of other entries from within the group. This was intended to increase the potential for variety (as opposed to repetition) in the statement set, anticipating that participants would view the statements of others and attempt to contribute an idea not already offered.

During the two-week brainstorming phase, study participants generated 901 statements about evaluation policy. During that time several recipients of the recruiting email contacted the researcher directly to say they would not be participating because they lived outside the United States. Several invited participants indicated they might not be participating because they felt they lacked adequate knowledge of the topic. The researcher encouraged all to participate nevertheless. After the specified period, the brainstorming phase ended.

The synthesis and reduction stage of concept mapping reduces the initial set of brainstormed statements generated in the brainstorming phase to a manageable number, of 100 or fewer, so that participants in the next stage of analysis can sort and rate the entire set without becoming discouraged and exiting before completion. In concept mapping projects where the number of distinct ideas is large, reduction is especially important. The challenge is to assure, through careful selection and editing, that this smaller set of ideas retains the full breadth and variety of meanings from the original statement set. The more successfully this is done, the richer and more meaningful the resulting concept map (Kane and Trochim 2007). This approach is similar to purposeful sampling strategies used in qualitative research known as "sampling for range", wherein the researcher identifies subcategories of a group being studied, and then gathers data from people in each sub-group (Small 2009; Weiss 1994).

The reduction process involves coding and categorizing, and then eliminating statements iteratively to obtain the most varied, though, not necessarily the most quantitatively representative statement set (Kane and Trochim 2007). So, for example, there may have been 50 statements on “evaluation use” and 10 statements on “institutional arrangements for evaluation” in the original statement set, for a 5:1 ratio. Yet in the final, reduced statement set, the ratio of statements on “use” to those on “institutional arrangements” might be 1:1, owing to many redundant “use” statements in the original set.

An Excel spreadsheet was used for reduction and synthesis. The researcher first read the complete set of statements and decoupled all compound statements so that each idea had its own row, increasing the number of statements to 1008. These were then coded interpretively, generating a word or phrase representing the main idea of the statement, following the line-by-line coding procedures developed by the constructivist grounded theorist Chamarz (2001).

While classical grounded theory holds that codes should “emerge” from the data in order to be considered truly inductive, constructivist grounded theory holds that data analysis is not a neutral act, and no researcher is a blank slate. Rather than to “discover” theory, the goal of grounded theory is to construct theory “based on what you discover is relevant in the actual worlds you study within this area” (ibid p. 335). The goal of the researcher is to approach the understandings of informants as closely as possible, but in doing so, the researcher actively engages in interpreting and constructing the interpretations and constructions of the those being researched. In the constructivist view, the “emergence” of theory depends partly on the conditions of research, the view of the researcher and the interaction of these with the understandings of informants. As such, different interpretations of the data beyond the one presented by a particular researcher are possible. A researcher’s reflexivity about the concepts and assumptions they bring to the work is crucial to its validity.

In approaching this data, the researcher brought background as a student of applied evaluation in a participatory tradition. The researcher's perspective was informed by perspectives from the systems evaluation literature, which holds that programs and policies are best understood when viewed together with and in relation to the contexts within which they function. The study is also predicated on certain assumptions, such as: high quality evaluation depends on a well-functioning evaluation system, and having such a system depends on having a comprehensive set of well-coordinated evaluation policies.

Prior to coding the data for this study, the researcher had read some (though not all) of the literature on evaluation policy. However, in coding the statements, frameworks from this literature were consciously ignored. Instead, the researcher strove to focus on patterns in the data. The researcher did have in mind, while coding, the Trochim definition of evaluation policy that prefaces the open-ended question to study participants, as well as the general structure of a comprehensive evaluation system, with mutually complementary parts working together. In this way, the ideas of "evaluation policy" and "evaluation system" serve as sensitizing concepts to guide the development of a clearer and better illustrated concept of evaluation policy in the U.S. federal context (Blumer 1954). Reading literature prior to coding accords with the Straussian view that grounded theory allows for a reading of the literature "as a stepping off point" to assist in the formation of questions to use in initial data collection (Rupsiene 2010, McCallin 2003, Strauss and Corbin 1998).

Each evaluation policy statement was assigned a thematic code (or "keyword" as it is called in the concept mapping literature), as well as a secondary code, using the same interpretive approach as above. Duplicate statements that were very close in wording and meaning to other statements in the set were removed. Thirteen non-responses were also removed from the set. These include single-word or partial-word strings that were obviously the result of a participant unintentionally hitting the enter key, and other statements that the researcher felt could not be considered a response to the focus prompt. These included, for example, "clearly the following lists all my concerns" and "can't think of anything".

In coding the statements, categories were developed iteratively, as in the constant comparison method of Glaser and Strauss (1967). As each statement was sorted into a category, it was compared with statements already in that category. When a statement didn't fit into any categories which had been created so far, a new one was generated. Categories were refined along the way. Once a first sorting of all statements was completed, the categories were then compared for overlap, differences in conceptual level, scope and type of category. Next, overarching themes were identified, and each statement assigned to one of these themes. An inclusion rule was then developed for each category. Along the way, the researcher iteratively tested proposed rules against each statement in a category, shifting statements and refining rules before finalizing the categorization scheme.

Next came reduction of the data. Weeding out statements not to be used in the next phase of analysis took several rounds of elimination. With each successive round, criteria to discard a statement relaxed from "an obvious duplicate statement exists in the current set" to "a very similar statement exists in the current set" to "any statement within the same sub-topic exists in the current set". In selecting statements for the final set, an effort was made to include at least one statement from each of the categories established in coding the original set of statements on evaluation policy. In doing so, care was taken not to privilege: 1) more general statements (principles); 2) more specific statements (rules) or 3) more forcefully worded statements. Care was also taken to include statements that reject study assumptions, such as the following: “There shall be no comprehensive set of evaluation policies.”

Last, grammar and spelling were corrected and any jargon rephrased so as to make each statement understandable to a general lay audience. For ease of reading, each statement was translated into a declarative form, so for example, the imperative: “In evaluation plans, explain stakeholder involvement in conceptualization of evaluation questions” became the declarative: “Evaluation plans shall include an explanation of stakeholder involvement in conceptualization of evaluation questions.”

Due to the large, nine to one ratio of the original statement set to the target size of 100 for the reduced statement set, this reduction process had to be repeated several times, and a large portion of the data discarded before the next step in the concept mapping analysis. To retain meanings that might otherwise have be lost, and in order to be able to compare discarded data to the data retained for the next stage, this process is fully documented in Appendices A, B, C and D. The 100 statements were randomized and uploaded to the Concept System’s website for participant sorting and rating. An invitation went out to potential sorters and a separate invitation went out to potential raters.

Sorting of statements was done online. To gather evaluators' input on how the 100 evaluation policy ideas should be grouped or categorized, a separate sample of 400 randomly selected members of AEA was invited via email by the researcher to participate in an online sorting exercise. (Interested brainstorm participants were also encouraged to participate in this phase of the study.) This phase took place in November of 2009, just before the AEA annual conference, when many invitees were occupied with preparing papers and presentations. Response to this invitation to sort was smaller than for the easier brainstorming task, with just 31 invitees beginning the sort task, and 28 completing. Sorting instructions asked participants to read the 100 statements, and then drag and drop each statement, one by one, to organize the set into groups by category, then give each group a label. Initially, this link was set up in a way that allowed sorters complete anonymity, but did not give them a way to stop part way through the task, save their inputs, and then reenter later. Several partial sorts came in and had to be discarded, possibly due to this arrangement. Halfway through the sorting period, the sign-in arrangements for this phase were adjusted to allow participants to create an anonymous userID so they could visit and revisit the link until done.

Rating of statements was also done online. For this task, instructions on the website first directed participants to read each of the 100 policy statements and then assign it a "merit" rating by clicking a button corresponding to their opinion about whether that policy idea should be implemented, with possible ratings ranging from "strongly disagree" to "strongly agree" At the end of the merit rating task, instructions asked participants to continue on to the "feasibility" rating task. Here, participants rated each statement on whether they viewed it as possible to implement, with possible ratings ranging from "very infeasible" to "very feasible".

The table below provides descriptive statistics on invitation samples and participation rates.

Table 3: Study response rates

Study phase	Total emails	Target participation	Started input	Completed input
Brainstorming	2000	100 (20%) for 300 statements	659	555 (28%) with 901 statements
Sort	400	25 (6.25%)	31	20 (5%)
Rate -"importance" -"feasibility"	427	 20 (5%) 20 (~5%)	 65 64	 63 (15%) 48 (~12%)
TOTAL	2,827			686 (24%)

This table shows that 2,827 emails were sent and lists response rates for each phase of the project. Participants who completed the brainstorming phase were encouraged to participate in subsequent phases. However, for all phases, online participation was anonymous, so it was not possible to track how many participants in the brainstorming phase also participated in sort or rate phases. 555 participants or 28% of those invited completed the brainstorming phase. Twenty participants, or 5% of those invited successfully completed the sorting phase. Of the 427 invited to participate in the rating phase, 63 raters, or 15% completed the "importance" rating and 48 raters, or 11% completed the "feasibility" rating. These response rates seem low, but are not unusual for web-based surveys. In a meta-analysis exploring factors associated with higher response rates in electronic surveys, Cook, Heath and Thompson (2001) report the mean response rate for the 68 surveys reported in 49 studies was 39.6% (SD=19.6%). While the average response rate for all phases of this study is within one standard deviation of the average at 24%, for certain phases, such as sorting, it falls well below. This is probably because of the large number of items (100), the high degree of complexity of the tasks, especially the sorting and rating tasks (one hour for sorting, 30-45 minutes for rating). Also, the sorting response rate may be lower than it would otherwise have been because the web site was initially set up not to allow participants to exit partway through the task and return to complete it later. Finally, the timing of the email solicitations to sort and rate, falling right before, during and immediately after the AEA 2009 annual conference, when many members were occupied with conference-related deadlines, activities and travel, may have dampened response rate somewhat.

To give an idea of how typical of the larger AEA membership the sample in this study is, demographic characteristics reported in the AEA 2007 membership survey are compared with the characteristics of all 601 participants in this study (including brainstormers, sorters and raters). AEA 2007 members are also compared with the 60 sorters and raters in this study, who completed more demanding and time-consuming tasks, and contributed more to the organization and interpretation of results. To test for the statistical significance of observed differences in the percentages of groups claiming specific characteristics, a one-sample Z-test of difference in proportions, with a critical value of .05 and two-tailed tests, is used.

In general, the participants in this study show a slightly smaller proportion of U.S. residents than the overall AEA membership, though the proportion of sorters and raters who were U.S. residents is not significantly different from the 87% of AEA members claiming a primary residence in this country. Compared with 29% of the overall AEA membership, a slightly larger proportion of study participants reported their primary work setting as a college or university. However, the proportion of study sorters and raters hailing from college or university work settings is not statistically different from that in the overall AEA membership. The proportion of study participants holding a doctorate is also not significantly different than the 52% of AEA members responding to the 2007 survey who reported having a doctorate. Comparisons of major work activity are difficult to make, since the categories of teaching, management and consulting were not used in the 2007 survey. The 9% non-response rate of study participants to these demographic questions, combined with the two-year gap between the 2007 AEA member survey and this study, make comparisons of the study sample of AEA members with the larger AEA membership less than precise, but provide a rough sense that the sample is comparable to the larger group.

Table 4: AEA Membership Compared with Study Participants

		% of 2007 AEA Membership (n=2,649)	% of study participants (n=601)	% of study sorters and raters (n=60)
Education	Doctorate	52	49	63
	Masters	41	37*	33
	Bachelors	7	5	1
	Other	0	1	1
	No response	0	9	2
		100	100	100
Work setting	College/university	29	36*	40
	School system	0	2	6
	State agency	4	4	1
	Federal agency	4	7*	8
	Local agency	4	2*	5
	Private business	16	17	18
	Non-profit	7	17*	15*
	Other	11	6*	5
	No response	25	9*	2*
		100	100	100
Major activity	Student	7	5	5
	Research	14	11*	6*
	Evaluation	49	36*	8*
	Teaching	-	9	8
	Management	-	15	23
	Consulting	-	11	45
	Other	15	3*	3*
	No response	15	10*	2*
		100	100	100
Country of residence	USA	87	80*	86
	Other	13	11	13
	No response	-	9	1
		100	100	100

*=difference in proportion is statistically significant at α=.05

Representation

After the sort and rate phase concluded, each participant's sort and rate responses were checked for completeness. Four incomplete participant sorts and four participant sorts with value labels such as "good policy" or "not feasible" were removed from the data. This left twenty complete sorts. Concept mapping processes uses multidimensional scaling and hierarchical cluster analysis to aggregate the sorting inputs of participants to create a map that could represent or stimulate discussion toward group consensus. While both multidimensional scaling and hierarchical cluster analysis can be performed in general purpose statistical programs, such as SAS and SPSS, this analysis used The Concept System software (Concept Systems Incorporated, 2005), which was designed to perform the sequence of analyses used in concept mapping. Sorting inputs were uploaded along with participant rating sets into the Concept Systems "core analysis" software for aggregation and the generation of concept maps.

The software quantitatively aggregates participant sorting data by first constructing for each participant whose sorting data is included an NxN binary, symmetric matrix of similarities, X_{ij}, where N is equal to the number of statements (N=100). For any two items i and j, if the two items were placed in the same pile by the participant, a one is placed in cell X_{ij}. Otherwise a zero is entered. Because in this study, 20 participants successfully completed the sorting, there were 20 different 100 X 100 binary similarity matrices. Next, a total NxN similarity matrix, T_{ij} is obtained by summing across the individual X_{ij} matrices. Thus, any cell in this matrix could take integer values between zero and the number of people who sorted the statements (in this case, 20). The value in each cell indicates the number of people who placed the i,j pair in the same pile.

So, for example, in this analysis, a cell in the total similarity matrix with a value close to 20 indicates that most participants whose data was included sorted these two statements together.

Through MDS, the aggregate binary symmetric similarity matrix becomes a two-dimensional map. MDS is a class of techniques that uses as inputs conceptual "proximities", such as psychological distance, similarity, relatedness, dependence, association, complementarity, or substitutability. For the concept map, a high value in any given cell of the aggregate similarity matrix translates to visual closeness on the two-dimensional map, meaning a short distance between the two points corresponding to these two statements. The configuration of objects on the MDS map is generated through an iterative process analogous to taking a table of distances between U.S. cities and then creating from them a map that best fits the proximities data. The software iteratively tests different possible dot map solutions before selecting the one with the best "fit" to the data. The output is a spatial representation consisting of a geometric configuration of points (Kruskal and Wish, 1978). While solutions with three or more dimensions are possible, the solution here is limited to two dimensions because, according to Kruskal and Wish: "…when an MDS configuration is desired primarily as the foundation on which to display clustering results, then a two-dimensional configuration is far more useful than one involving three or more dimensions" (p. 58).

The Concept Systems software performs hierarchical cluster analysis on the x,y coordinates from the point map using Ward's algorithm (Everitt, 1980). Ward's algorithm is one of a family of cluster analysis formulations used for forming hierarchical groups of mutually exclusive subsets. The objective function for Ward's algorithm is to minimize the total distance between statements within each of the clusters. The process begins with one statement per cluster, then moves to progressively fewer clusters, until the entire set is one large cluster. Given n sets, the procedure permits their reduction to n-1 mutually exclusive sets by considering the union of all possible n(n-1)/2 pairs and selecting a union having a maximal value for the objective function. Ward's algorithm was intended for use in classifying, e.g. making taxonomies of plants and animals and for organizing and cataloguing materials so as to facilitate retrieval of information (Ward 1963). While both multidimensional scaling and hierarchical cluster analysis can be performed in general purpose statistical programs, such as SAS and SPSS, this analysis used The Concept System software (Concept Systems Incorporated, 2005), which was designed to perform the sequence of analyses used in concept mapping. The software iteratively tests different possible dot map solutions before selecting the one with the best "fit" to the data.

Interpretation and further analysis

Typically, cluster analysis in concept mapping begins with a many-cluster solution and then examines successively fewer-cluster solutions, making an interpretive judgment each time about whether the merging of clusters results in a configuration that is meaningful and interpretable. (Kane and Trochim, 2007). This may be done in collaboration with study participants or by the researcher alone. Here, the researcher worked in consultation with another researcher familiar with the concept mapping method, examining solutions beginning with a twenty-cluster solution and ending with a nine-cluster solution. The researcher gave a name or label to each cluster to summarize and represent the statements inside. The content of each cluster was compared with the pre-reduction statement set from prior stages of analysis, and with the categories in these results with the Trochim (2009) evaluation policy taxonomy.

To examine the conceptual coherence of each cluster on the concept map, and to understand better the conceptual connections among clusters in the eyes of study participants, a bridging and anchoring analysis was conducted, showing which statements and clusters on the concept map were also sorted with statements distant on the map by some participants. To illuminate the interrelationships among clusters and regions as seen by participants, a spanning analysis of the statements on the map was conducted, to show which distant ideas they were sorted with, and what these connections might reveal about sorters' individual categorization schemes.

To examine variation in the rating responses in this study, participant demographic data was combined with participant rating data in a "pattern match" graph. To generate these results, the analysis averages participant ratings for each statement, and then averages these average statement rating values across statements within each cluster, for an average cluster rating. For each scale, average cluster ratings are presented on a vertical axis, with the lowest values in the range of responses for that scale at bottom and highest values in the range at top. The two vertical axes cluster average ratings are presented side by side, appearing as a "ladder" display. The relationship between the respective rating scales for each cluster appear as the rungs of the ladder. The more smaller the slope a rung, the more similar the relative average ratings for that cluster across the two scales. The steeper the slope of the rung, the more different the relative average ratings for that cluster across the two scales. However, differences in absolute average cluster rating values may be due to tendencies to rate higher or lower within different participant groups, or to different interpretations of scales within those groups, the comparison of relative average cluster rankings is considered more meaningful. The analysis computes the correlation of the two sets of average cluster ratings and generates an r value, corresponding to the standard Pearson product moment correlation coefficient, ranging between -1 and 1. Values closer to zero mean the two sets of cluster rankings are not strongly correlated, and values closer to the absolute value of one mean the two sets of cluster rankings are strongly correlated.

To identify the evaluation policy ideas viewed by study participants as both highly worthwhile and highly do-able, the rating data from the "importance" and "feasibility" rating scales was combined in a "go-zone" analysis. According to Kane and Trochim (2007) "A go-zone is a specific type of bivariate plot of the data in a pattern match, generally showing the averages for each statement within a cluster. It plots the statement results in an x-y graph, divided in quadrants above and below the mean value within the cluster of each rating variable. The term 'go-zone' springs from the fact that upper-right quadrant displays statements of a cluster that were rated above average on both variables. In many situations, these will represent the most actionable statements within the cluster." (p.22)

Limitations

The brainstorming sample, when asked for suggested evaluation policies, returned a very high response rate for web-based surveys, resulting in an unexpectedly large set of data. The response rate for raters was lower, but still within the typical range for web-based surveys. Those asked to complete the sorting task responded at an even lower rate. This was probably due to the complexity of the task, combined for anonymous inputting, which made it impossible for sorters to leave an incomplete, then return and finish later. Partway through the sorting period, several incomplete sorts were discovered, and these had to be discarded. The log-in for the sorting webpage was then changed to allow users to specify an anonymous ID they could use to return repeatedly. Apparently, though, this change was not made in time to increase the sort participation rate substantially. The study was left just twenty complete sorts, a small number for sorting in this concept mapping method.

While the American Evaluation Association includes a large and diverse membership, the results of this study are limited to the accessible members of that organization, and should not be taken as representative of the thinking of the membership of the American Evaluation Association.

A further limitation of the study is that it does not include input from other professional organizations with members engaged in evaluation or evaluation research within or paid for by the federal government. These include, most notably, the Association for Public Policy and Management. According to AEA's 2007 member survey, AEA's membership is primarily (49%) engaged in practical evaluation in non-academic settings, with 15% reporting they are college or university faculty members or instructors teaching evaluation, 14% reporting they are evaluation researchers and 7% reporting they are students of evaluation. By contrast, APPAM's website stated in 2009 that its membership was approximately 70% academic (including students), and 30% non-academic, with most non-academic members employed at policy research organizations and the public sector. The educational background of AEA members is diverse, but rests predominantly in education, psychology and sociology. Many of its members work in field settings where practical constraints dictate quasi-experimental and non-experimental designs. By contrast, APPAM's membership is known for an emphasis on economic analytical approaches and the use of experimental methods in policy analysis, including evaluation.

A further limitation of the study is that it does not incorporate input from the other main stakeholder groups with an interest in evaluation policy, namely decision makers who are consumers of evaluation results, nor the general U.S. public, who are the beneficiaries of policies and programs in this country.

For the reasons above, the statements in these data cannot be taken as universally representative of evaluation practitioner thinking on the subject of evaluation policy, nor of the thinking of all American Evaluation Association Members, and are more properly taken as informative and enriching of a large and expanding set of researcher and theorist inputs into the classification of evaluation policy.

CHAPTER FOUR: RESULTS

This chapter presents the results of the concept mapping data collection and analysis, including brainstormed participant survey responses, researcher reduction and synthesis results, participant sorting and rating results, with analyses combining these different sets of results to form the basis of a taxonomy with classification criteria.

Brainstorming responses

Response by AEA members receiving the survey invitation to brainstorm evaluation policy ideas was quick and fairly large for online surveys, with 574 invitees beginning and 554 successfully completing the brainstorming task, entering a total of 920 statements into the online survey system within the first three days. For the original set of responses, please see Appendix A. What follows is an impressionistic overview of the initial data, immediately followed by a more precise description of coding and categories, as well as synthesis and reduction results.

Most of the statements in this original, larger set of 920 responses speak to one of four action areas: 1) values that should guide evaluation practice; 2) setting up an evaluation system (in the sense of setting up organizational rules and roles, building the capacity of staff to conduct evaluation, and arranging for oversight of evaluation activities); 3) conducting evaluation, in the sense of designing an evaluation, selecting methods and measures, collecting and analyzing data and 4) communicating about evaluation results and making sure they are used.

The data include policy ideas ranging from very specific methodological rules, such as "Evaluations should be held to obtaining an 80% response rate" (statement 543) to more general principles based in values, such as “Try to understand that evaluation is an intervention, and make the change in a positive direction” (statement 810). Most of these ideas do not specify the organizational level at which they are to be implemented, and only a few directly refer to a specific organizational body. For example statement 765 designates an evaluation point person within each federal agency to serve on a central coordinating committee. Only a few could be considered high-level “meta policy”, in the sense that they articulate policy about policy, for example, statement 778, which calls for Congressional guidelines about what is and is not properly in the scope of evaluation policy.

The set of responses from AEA members reflect some of the tensions at the heart of the evaluation profession that related to evaluation field’s historical relationship with the governmental system in which it is embedded. Statements calling for the tailoring evaluation approaches to stakeholder needs, the features of a program, its context, and target population stand in tension with the statements relating to standardization of practice, in some cases for better collaboration across program and agency boundaries. On the other hand, statements calling for the involvement of stakeholders in evaluation processes and integration of planning and evaluation functions stand in tension with statements outlining policies that would protect evaluator independence from outside influences on their practice. Statements calling for the use of external, independent evaluation experts stand in tension with statements calling for evaluation capacity building for more and better internal evaluation by program managers and staff.

Not surprisingly, statements in this group of evaluation policy ideas from American Evaluation Association members reflect concerns for the professional self-interest of evaluators, such as statements about using only certified or properly qualified professional evaluators, or involving professional evaluators in the design of evaluation requirements for programs and decisions about who gets funded to do evaluation. However, study participants are also clearly concerned with the values and sense of higher purpose they feel should inform evaluation practice, as reflected in statements about social justice in evaluation practice, the ethical treatment of subjects, transparency in evaluation, respecting minority voices and understanding cultural differences.

Statement synthesis and reduction

Concept mapping calls for reduction of the set of brainstormed statements generated by participants to a set of 100 or less, for ease of sorting in later stages. In concept mapping studies where the size of the original set of brainstormed statements is small, synthesis and reduction decisions are less important for the ultimate concept mapping results. However, because of the high ratio of discarded to retained data here, decisions about what to keep and what to discard are potentially more influential. Many of the discarded statements are simply identical to or very similar to statements that were selected to appear in the reduced, final set used for sorting and rating. However, some discarded data may contain meanings that are lost in the reduction process.

What follows is a detailed description of the synthesis and reduction process for this study, along with a comparison of the original set of evaluation policy ideas with the final, reduced dataset that became the input for later rating and sorting phases. A chart outlining the reduction process appears below, followed by a description of that process.

Table 5: Synthesis and reduction process

	Original statement set (N=920)	**After splitting compound statements (N=1008)**	**After removing non-responses (N=995)**	**After synthesis & reduction (N=100)**
Decision rule	ALL saved inputs	Split if two or more ideas; end up with one statement per idea.	Discard if doesn't answer the prompt or can't work with answer.	Discard if redundant, or if similar idea is retained in final set; sample for range
Appendix	A original statement set with all statements, no splits	B List of compound statements with break-out statements	C list of removed, non-response statements	D final statements by super category with original categories E final statement set

The original data contained 920 statements from 554 brainstormers. The first step in preparing the statements for rating and sorting was to separate each compound statement into two or more distinct ideas, leaving one idea per statement. This process identified 51 statements with more than one idea in them. Splitting compound statements into single-idea statements increased the size of the statement set by 88, to 1008. (For the list of compound statements showing how these were split into single-idea statements, please see Appendix B.) The next step was to remove 13 responses that did not seem to address the focus prompt at all, bringing the statement set down to 995 distinct ideas. (For the list of statements discarded as "non-responses", please see Appendix C.) Initial coding of these ideas organized the set into small, fine-grained categories. These finer-grained categories were then reviewed and many commonalities and overlaps across categories found. As a result, the 77 categories were combined into 17 larger “super categories” under the names of Approach, Method, Evidence, Standards, Tailoring, Transparency, Values, Resources, Capacity Building, Coordination, Integrate evaluation, Independence, Who evaluates, Scope, Reporting, Uses of Evaluation and Meta-policy. For each super category, ideas were discarded in several iterations until all that was left were the 100 statements selected to represent each distinct “super-category” in the final, reduced set. Appendix D contains a set of tables organized by the 17 “super categories” showing the final set of 100 selected statements mapped to the initial 77, finer-grained categories from which they were chosen, along with the statement numbers of discarded statements in those categories. Appendix E contains the final set of 100 statements, showing each statement’s original ID number (1-920) at right, and its new ID number (1-100) at left.

Comparing the original statement set to the final, reduced set

In general, the reduced set represents every major category of the original set, as shown in the grids matching selected statements back to categories from which they were chosen in Appendix D. One exception is the finer-grained category “Purposes of evaluation”, which includes statements about the various possible purposes of evaluation and various possible uses of evaluation results. Purposes mentioned in these statements include: the purpose of evaluation being to judge the merit of a program (a dimension central to the classic definition of evaluation in the field), as in statement 537; accountability (e.g. statement 488), oversight (e.g. statement 737), program management (e.g. statement 241), quality improvement (e.g. statement 536), providing “useful information to the primary intended users” (see statement 903), learning what works for future funding decisions (e.g. statement 855) informing decision making by policy makers (statement 914), tracking the long-term impact of programs (statement 279) and “adding to the body of research on evaluation” (statement 113). The only statement selected to represent in the reduced set this rich range of ideas on the purpose of evaluation was: “Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight” (statement 737).

In the effort to include all distinct and unique ideas and avoid the temptation to focus on the most frequently occurring ideas, this synthesis and reduction process sometimes left out frequently occurring ideas entirely in favor of unique ideas at extreme ends of a conceptual range. For example, the ideas selected from the “Methods” super category prescribe two almost opposite methodologies--developmental approaches (a combination of statements 643 and 17) and random experimental design (statement 774). While these two unique statements represent the two extreme ends of a conceptual range, the center of the range, with more moderate statements calling for high-quality or rigorous methods implemented by experts (statements 24 and 698), is not represented in the reduced set at all. Another example of this: left out of the reduced set was an oft-repeated idea that, in general, evaluation should be integrated at the earliest stages of program development (see, for example, statements 31, 68 and 91). However, the final set includes another, more unique and specific policy idea on this topic: “Evaluators shall serve as key members of the planning body for each project and program (statement 363). A final example is in the “Resources” category, although the original dataset contains several statements suggesting establishment of a universal, fixed minimum of 10-15% of project funding dedicated to evaluation (see for example statements 65, 132 and 340), this particular idea does not appear in the reduced set. Instead, a more unique statement about setting guidelines for funding of evaluations appears (see statement 831).

As stated earlier, the original intent was to sacrifice quantitative representativeness in favor of range. To compare category proportions in the original, larger set of statements with those in the reduced, closely related super categories were first combined into eight larger groupings. Approach, Method, Evidence and Tailoring combined to form a larger grouping called Approach, Method and Evidence. Transparency and Values combined to form a larger grouping called Values. Resources, Capacity Building and Coordination combined to form a larger grouping called Institutionalizing evaluation. Integrate evaluation, Independence, and Who Evaluates combined to form a larger grouping called Roles and Relations. Reporting and Uses of evaluation combined to form a larger grouping called Reporting and Use. Standards, Scope, and Meta-policy remain stand-alone super categories. The proportional make-up of original and reduced statement sets as organized by these eight larger groupings are shows in Figure 2 below.

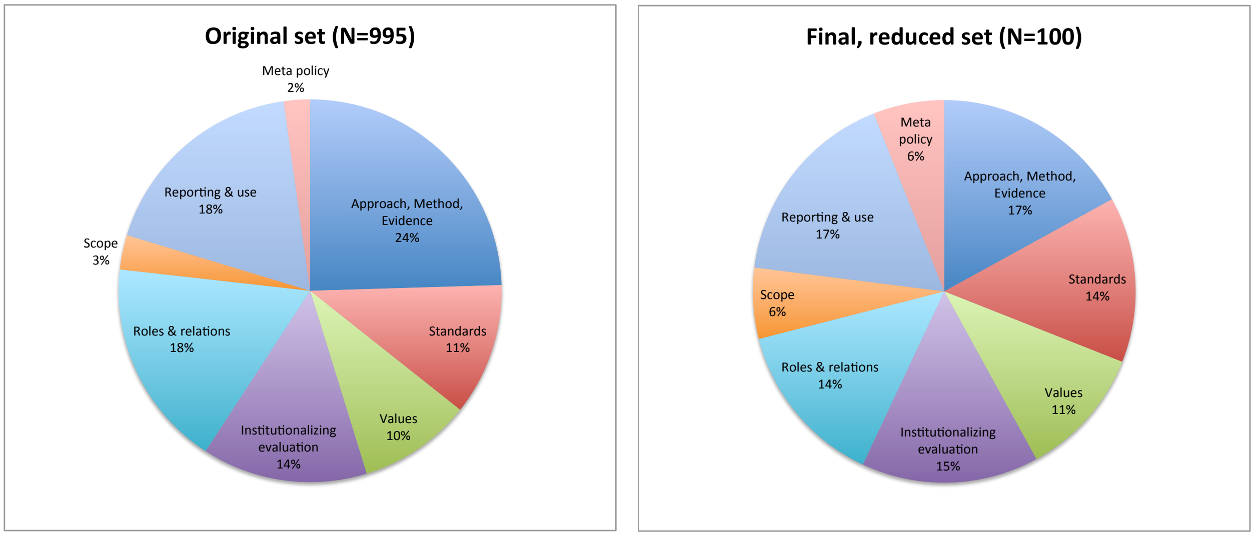

Figure 2: Original and reduced statement sets, category proportions

The reduced statement set is slightly more balanced by proportion across these larger groupings. Specifically, the reduced set contains a smaller percentage of statements from “evaluation “approaches, methods and evidence”, on evaluation “roles and relations” and on evaluation reporting and use of evaluation results, but a larger percentage of statements on standards for and values in evaluation, on institutionalizing evaluation, on evaluation scope and on evaluation meta-policy.

Concept map

The instructions to sorters in this study urged them to avoid grouping them according to perceived value or feasibility of the ideas. However, eight sets of participant sorts containing categories such as “bad idea” or “good idea had to be discarded, leaving twenty. These were added together and an MDS process applied to generate the point map below.

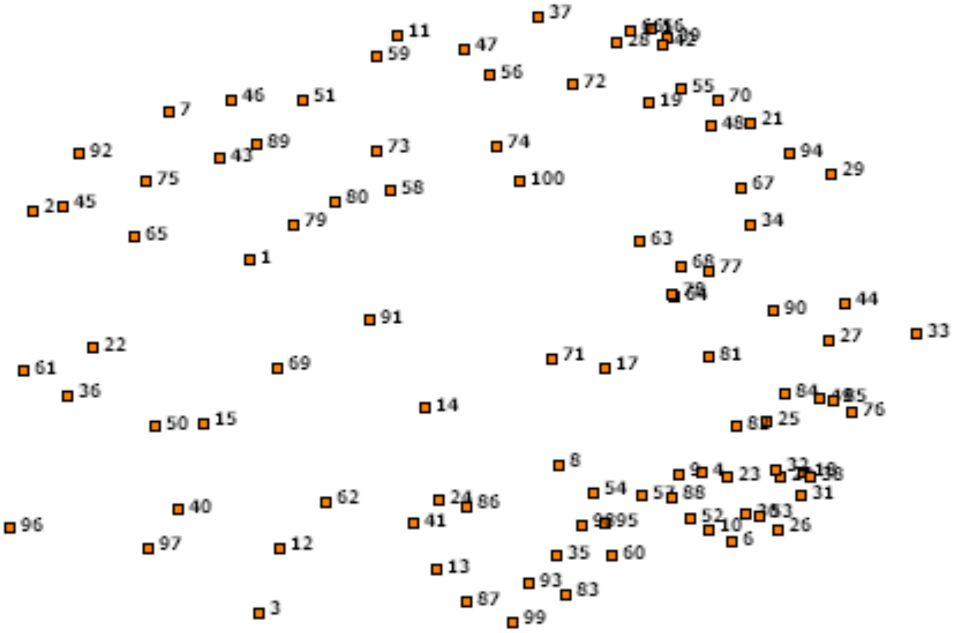

Figure 3: Point map

To assess the goodness-of-fit of the best-fit point map produced here with the distance values in the aggregate similarity matrix, the MDS analysis produces a diagnostic called a “stress value”. “Stress” corresponds roughly to the portion of the variance explainable by MDS. If the stress value is low, the fit of the map is better. A high stress value can imply the map is not a good representation of the sort data. According to Kane and Trochim (2007), higher stress values in concept mapping may mean “there is more complexity in the similarity matrix than can be represented well in two dimensions, that there was considerable variability in the way people grouped the statements, or both” (p. 98). Stress values will be better when there are more than 25 people sorting (Rosas, 2011). Yet despite the relatively small number of sorters (20), the stress value for this study was .27794 after 9 iterations, indicating slightly below the average stress for smaller concept mapping studies (Trochim 1993). Given the small number of sorters in here, this value is a bit surprising. This could be interpreted as an indication of exceptional homogeneity within the small group of sorters. However, it is important to note that according to Trochim, stress is “highly sensitive to slightly movements in statements on a map not likely to have any meaningful interpretive value in concept mapping, so stress can only be considered a rough measure of the fit of the map” (Kane and Trochim 2007, p.98).

Stress estimates the deviation of the concept map from the aggregate sort matrix. What it cannot show is the degree to which individual sort inputs differ from the “typical” sort expressed by the aggregate matrix. For a sense of the variation hidden within the aggregate solution, an “individual-to-total” reliability analysis was conducted (Trochim 1993; see also Rosas 2005). To estimate this coefficient, which is analogous to average inter-rater reliability, each participant’s individual binary sort matrix was correlated with the total similarity matrix. The 20 correlations were averaged, and the Spearman-Brown correction for differences in sample sizes was applied. The average inter-rater reliability estimate for the participant sort inputs was 0.88 (*df* = 5049, *p* < .01), indicating statistical consistency in the sorted relationships across participants. This value is above the 0.86 average reliability for thirty-three small N concept mapping studies found in a Trochim (1993) pooled study of 38 concept mapping projects with an average of fifteen sorters. However, it is below the 0.91 average reliability found for 69 concept mapping projects in a concept mapping pooled study with an average of 24 sorters (Rosas, 2011).

In starting to look for groupings of policy ideas here, note that on this first concept map, points in the lower right region appear more closely clustered than those on the left side of the graph. A handful of statements on the upper right are also set close together. An intuitive interpretation might be that these more closely spaced groups represent groupings of more closely related policy ideas. Following is a systematic investigation of the conceptual connections among statements as represented in the concept map.

Bridging and anchoring statements

It is important to note here that MDS may place a statement in a particular spot on the graph because many people sorted it with the statements appearing nearby. Such a statement is considered an "anchoring statement". On the other hand, MDS may place another statement in a particular spot because it was sorted by different participants with more distant statements on either side. In this case, the software positions the statement point at an intermediate position between the more distant statements. A statement sorted in this way is considered a "bridging" statement.

Few statements have only “anchoring” or only “bridging” connections to other statements on the map--most have a mix. To show whether a statement has mostly bridging or mostly anchoring connections, the Concept System assigns each statement a “bridging value”. Bridging values range from 0 to 1.0, with 0 as the least bridging (or most anchoring) and 1.0 as the most bridging (or least anchoring). A point bridging map appears in Figure 4 below.

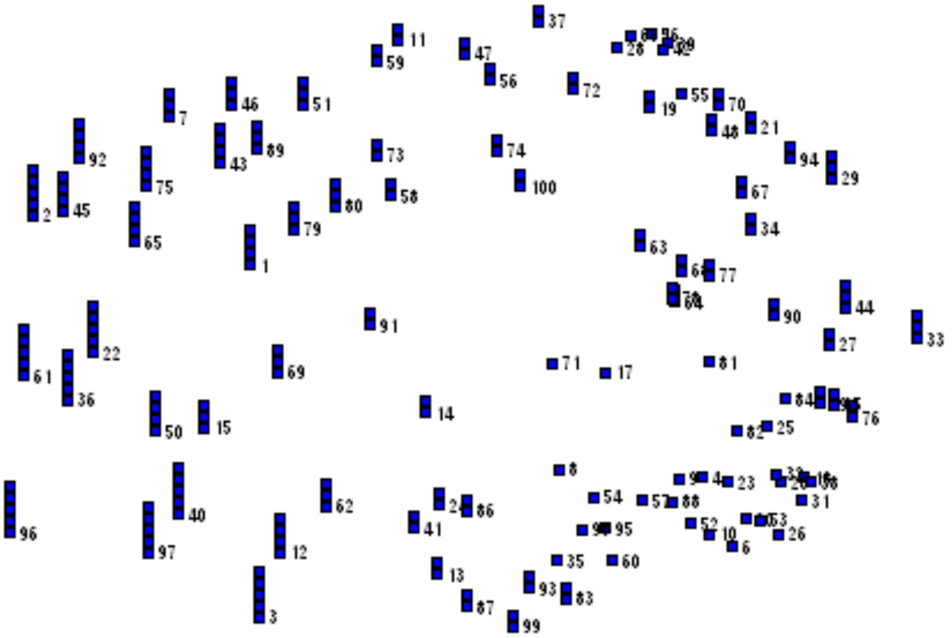

Figure 4: Point bridging map

On this map, statements with a higher “bridging” value appear as taller stacks of dots, and statements with lower “bridging” values (“anchoring” statements) appear as lower stacks of dots. In keeping with the intuition that the closely-spaced points in the lower right section of the map are closer conceptually, this more closely-spaced area of the map does, in fact, contain more "anchoring" statements. The pattern of "bridging" and "anchoring" statements on the concept map becomes more meaningful after statements are grouped in meaningful clusters.

Cluster analysis

The next step in concept mapping is hierarchical cluster analysis, in which different possible groupings of the ideas on the concept map, are examined. Cluster solutions ranging from nine to twenty-three clusters were examined. All possible cluster solutions with fewer than twenty clusters generated combinations of statements that, in the view of the researcher, did not seem to hang together conceptually. To achieve the greatest possible intra-cluster coherence, this analysis uses a twenty-cluster solution, which is a relatively high number of clusters for a concept mapping project. The researcher gave a name or label to each cluster summarizing the statements inside. The cluster solution is shown in Figure 5 below. (For the set of 100 evaluation policy statements list organized by cluster, please see Appendix F.)

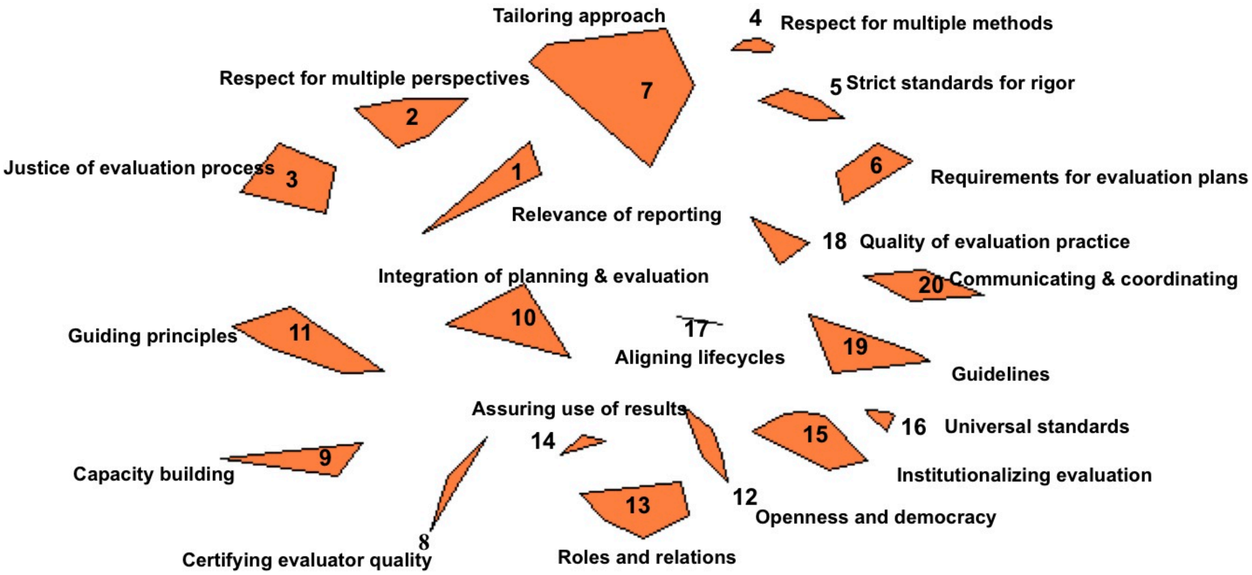
Tailoring approach
4 Respect for multiple methods
Respect for multiple perspectives
7
2
5 Strict standards for rigor
Justice of evaluation process
3
1
6
Requirements for evaluation plans
Relevance of reporting
18 Quality of evaluation practice
Integration of planning & evaluation
20 Communicating & coordinating
10
Guiding principles
11
17
Aligning lifecycles
19
Guidelines
Assuring use of results
14
15
16 Universal standards
Institutionalizing evaluation
Capacity building
9
13
12
Openness and democracy
8
Certifying evaluator quality
Roles and relations

Figure 5: Cluster map

In general, each of these twenty clusters has at least some degree of face value conceptual coherence. However, some clusters center on a particular theme or function, while others are unified by the similar form or level of the policies inside. This suggests study participants were not 100% consistent in following study instructions for forming categories, and could undermine the value of these results as a basis for a taxonomy with non-overlapping categories. Following is a description of the contents of each cluster.

Cluster 1 "*Relevance of reporting*". These statements all touch on evaluation results, and assuring results are reported and used for various purposes, including learning, "to benefit the population or program" or to inform policy.

Cluster 2 "*Respect for multiple persp*ectives". These ideas all correspond to the planning stage of an evaluation, including setting up an evaluation team, researching the literature, and specifying the level of privacy for human subjects, surfacing assumptions in a logic model, conceptualization of program goals. These statements are about inputs into the evaluation, such as stakeholders, human subject data, evaluation expertise, best practices literature, and underlying assumptions.

Cluster 3 "*Justice of evaluation process*". Several statements in this cluster are about upholding the goal of social justice in how evaluation is conducted. Two statements address the need to bring out or attend to minority views or the impact of a program on minority groups.

Cluster 4 "*Respect for multiple methods*". Five of six of these statements have to do with matching method to situation. The sixth calls for measuring all outcomes, not just expected ones. The unifying theme is being flexible and not rigid in methodological approach to evaluation.

Cluster 5 "*Strict standards for rigor*". These statements seem closely related to cluster 4, but in contrast to the flexibility theme of 4, these are more rigid requirements for doing evaluation. It appears these statements may have been sorted together because most respondents disagree with them, or because they correspond to typical, conservative, rigid grant requirements.

Cluster 6 "*Requirements for evaluation plans*." Within the topic of evaluation plans, these statements vary in rigidity (setting the plan at the outset and not varying from it) versus flexibility (leaving open the possibility of changing the plan mid-stream if needed). The cluster includes a statement about process evaluation, which may be related to the others here because process evaluation can examine whether an evaluation plan was implemented or changed mid-stream.

Cluster 7 "*Tailoring approach*". Several statements in this cluster have to do with respecting and incorporating stakeholder perspectives and local cultural perspective in an evaluation. Others have to do with tailoring the evaluation to the situation or intended uses of the evaluation. A unifying goal of all these statements could be full utilization of evaluation results.

Cluster 8 "*Certifying evaluator quality*". Two statements in this tiny cluster relate to certifying or rating evaluators. A third statement here calls for publishing evaluation results to identify best practices. The common purpose of these suggested policies might be inform consumers of evaluation services or to increase demand for those with evaluation credentials and experience.

Cluster 9 "*Capacity building*." The statements in this very small cluster relate to training, technical assistance and "evaluation culture", a term usually used to refer to attitudes about evaluation or willingness to do evaluation among personnel in an organization. All three policy ideas relate to building the capacity of organizations (evaluators) to do high quality evaluation.

Cluster 10 "*Integration of planning and evaluation*." The statements in this tiny cluster call for considering how results will be used when designing an evaluation plan, including evaluators in planning and designing a program, and using data for the purposes intended in the evaluation plan.

Cluster 11 "*Guiding principles*". These statements refer to values that should guide the way evaluation is done. Specific standards mentioned include 1) the AEA Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare and 2) the standards for evaluation developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy. Other statements call for "valuing" evaluation results in alternative ways.

Cluster 12 "*Openness and democracy*". In this cluster, two statements have to do with transparency of evaluation plans, findings and costs. Two statements have to do with opening up for wider competition federal grant funding for evaluation. A fifth has to do with assuring follow-up evaluation on federal grants, opening up grantees to ongoing (rather than just one-time) scrutiny.

Cluster 13 "*Roles and relations*". Proposed policies here seek to assure quality by more clearly articulating and supporting those in the role of evaluator, and by clarifying relationships between evaluators and stakeholders.

Cluster 14 "*Assuring use of results*". The proposed policies in this cluster aim toward fuller use of evaluation results, articulating when and how results are to be used.

Cluster 15 "*Institutionalizing evaluation*". This large and varied cluster includes policies establishing institutional structures for the conduct and oversight of evaluation activities across government. Topics include personnel structures for oversight of evaluation, definitions of success (in grants and performance indicators), funding structures for evaluation (as distinct from program funding), and the overall capacity of the federal evaluation workforce. One shared purpose of many of the policies in this cluster seems to be to protect high quality evaluation against political influences.

Cluster 16 "*Universal standards*". Two of these statements relate to a standard language for evaluation across the whole government. Two of them relate to a core set of evaluation policies across the whole government, at all levels. One seems unrelated to either of these themes, having to do with requiring an evaluation plan for all federally funded projects and programs. While these statements may be closely related to those in cluster 15, the unifying theme joining them seems to be the integration of policies across the entire government. This cluster contains the statement that expresses the central assumption of this study: "There shall be a core set of evaluation policies that apply to all federal evaluations."

Cluster 17 "*Aligning lifecycles*." This tiny cluster contains one statement about allowing new programs time to develop, or progress further through their lifecycle, before requiring them to report impact, as would be appropriate for a more mature program. The second calls for efforts to decrease "pro-forma" evaluation, which may refer to evaluations manipulated in response to unreasonable government requirements to show change before any change attributable to the program could have had a chance to take effect. The term "aligning lifecycles" refers to allowing a program early in its development to use an evaluation approach suited to programs in this early stage, for example implementation evaluation.

Cluster 18 "*Quality of evaluation practice*". Statements in this cluster are quite varied, including requirements for transparency around how data is analyzed, criteria for quality of innovative practices, and follow-up evaluations. One statement asserts the principle of respecting local, contextual realities as the standard for truth (rather than generalizability), which seems to encourage flexibility. Another requires that only the English language be used in evaluation studies, a more rigid criterion for quality.

Cluster 19 "*Guidelines*". This cluster is also quite diverse. Several statements use the word "guidelines", calling for a broad framework to cover one or another aspect of evaluation within government, such as when to evaluate, and at what level to fund evaluation, and how "R&D" (research and development) shall be handled, as distinct from evaluation. This cluster seems to be unified not by a topic but by the form or level of policies. Most of these policies involve empowering some other level of the organization to make more specific policy, and rest somewhere in between the most abstract level of the principle and the most specific level of a prescriptive rule. Embodying this is a statement opposing the central assumption of this study: "There shall be no comprehensive set of evaluation policies".

Cluster 20 "*Communicating and coordinating*". Evaluation policy ideas here call for a clearinghouse of evaluation methods for use across the organization, using a shared set of guidelines for standards of evidence and even identical measures across programs and agencies. The common goal of these policies is coordination across sub-units of the organization.

Cluster bridging and anchoring results

One possible application of the clusters of statements in this next map is as a set of non-overlapping evaluation policy types. In this application, the usefulness of the results rests on the distinctness and non-overlapping quality of clusters. Unfortunately, the initial point map that shows some regions contain more closely-spaced groups of ideas than others, suggesting not all clusters are equal in their degree of cohesiveness. Why?

Recall that on this map, statements appearing near each other may do so because many participants sorted them together. Other statements may appear together because there was lack of consensus among sorters about where to put them, so the software placed them at a midpoint between two or more other statements with which they had been sorted. If a cluster is composed mostly of statements that are highly “anchoring”, they are conceptually more related to one another, in the eyes of participants. If a cluster contains many bridging statements and few anchoring ones, this suggests an interpretation of the map other than as a set of non-overlapping categories may be in order. To examine more closely the conceptual coherence of each cluster as viewed by participants, a cluster bridging map was generated. Figure 6, below, displays clusters in stacks whose height corresponds with statement bridging values averaged across the cluster. A cluster with more layers stacked up is one whose statements have a higher average bridging value.

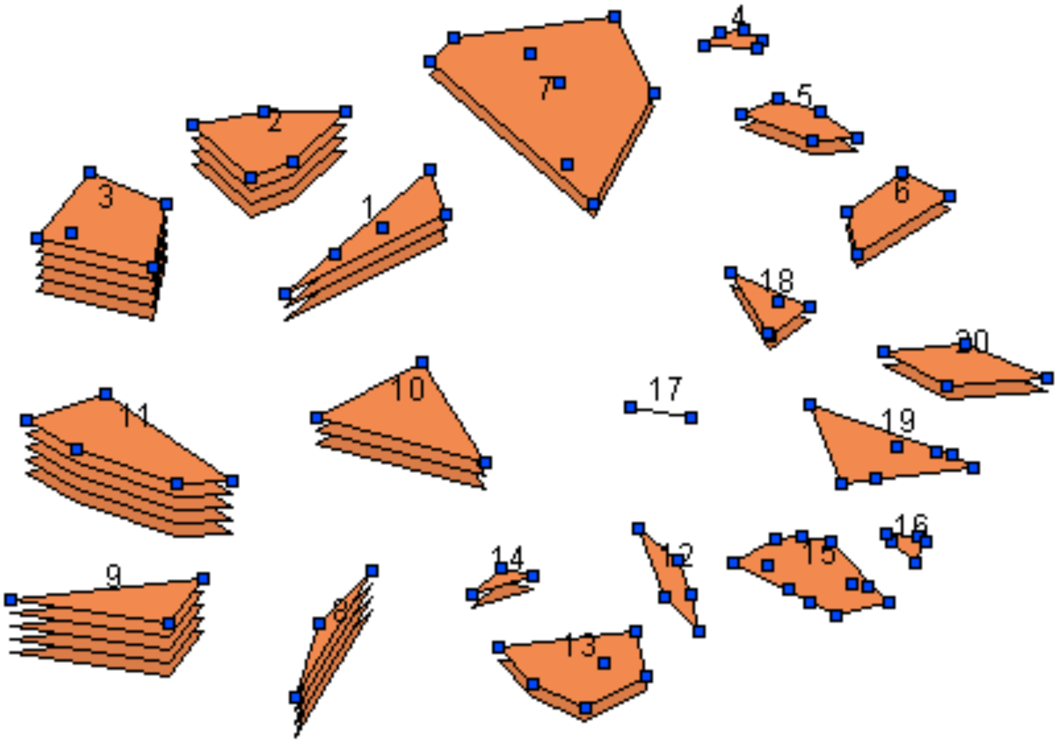

Figure 6: Cluster bridging map

This map shows that clusters in the lower right area of the map, including cluster 16 “universal standards”, cluster 15 “institutionalizing evaluation”, cluster 4 “respect for multiple methods”, cluster 12 “openness and democracy”, cluster 17 “aligning lifecycles”, cluster 19 “guidelines” and cluster 5 “strict standards for rigor” all have very low average bridging values (between 0.05 and 0.24). The low average bridging values suggest these clusters may be conceptually more isolated and therefore more cohesive in the eyes of study participants as a group. A closer look at policies in these clusters reveals they contain a high concentration of more specific rules. There may be more agreement among participants about where to sort these policies because their content is more specific and isolated in scope.

Clusters in the middle, middle top and middle bottom, including clusters 18 “quality of evaluation practice”, 14 “assuring use of results”, 13 “roles and relations”, 7 “tailoring approach”, 6 “requirements for evaluation plans”, 20 “communicating and coordinating”, 1 “relevance of reporting” and 10 “integration of planning and evaluation”, have somewhat higher bridging values that could still be considered low (between 0.26 and 0.44). These clusters contain a mix of bridging and anchoring statements. A closer look reveals the content of these clusters tends to be in between a general principle and a very specific rule.

Clusters at the far left side of the map, including cluster 2” respect for multiple perspectives”, cluster 8 “certifying evaluator quality”, cluster 3 “justice of evaluation process”, cluster 11 “guiding principles” and cluster 9 “capacity building” have moderate to high average bridging values (between 0.56 and 0.88). The low average bridging value of these clusters could indicate they are less conceptually coherent in the eyes of study participants as a group. A closer look suggests these clusters, especially 2,3 and 11, contain a higher concentration of general principles, which are likely to have conceptual connections to several other, more specific policies. So it makes sense that participants would see them as having more connections to many other statements on the concept map.

Spanning analysis

To examine more deeply the conceptual relationships among the 100 suggested evaluation policies, as seen by participants, several highly bridging "interloper" statements on the map were examined. This analysis shows lines spanning the distance from a selected statement to other statements with which participants sorted it. Thicker lines mean more participants sorted the two statements together. An example of a highly anchoring statement is shown below.

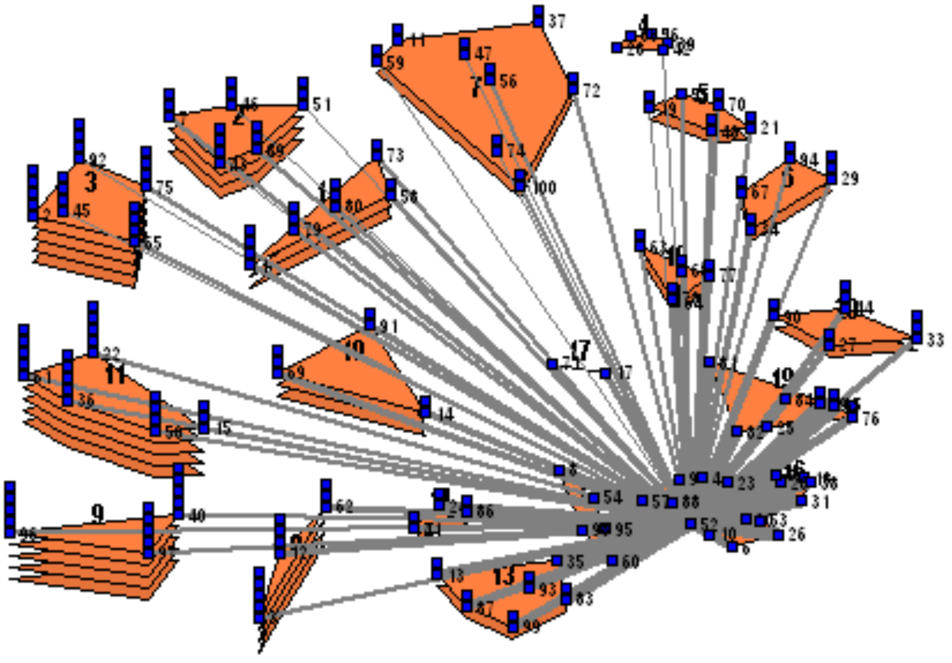

Figure 7: Spanning analysis for statement 52

Statement 52, “A separate evaluation contract shall be included in major grants to ensure that funds available for evaluation are not reallocated to program efforts”, has the lowest possible bridging value of 0.00. This value is so low partly because the statement has many moderate to strong ties to statements in its own cluster of “Institutionalizing evaluation” (with 8 or more people sorting statement 52 with statements in its own cluster) and to statements in closely neighboring clusters especially clusters 12 “Openness and democracy” and 19 “Universal standards”.   However, all its ties to statements in more distant clusters are weak ones, with 3 or fewer people sorting 52 together with these more distant statements. Similarly, the more loosely-spaced points at the left of the map that appear conceptually more distant from one another are, in fact, mostly “bridging” statements. An example of a highly bridging statement is shown below.

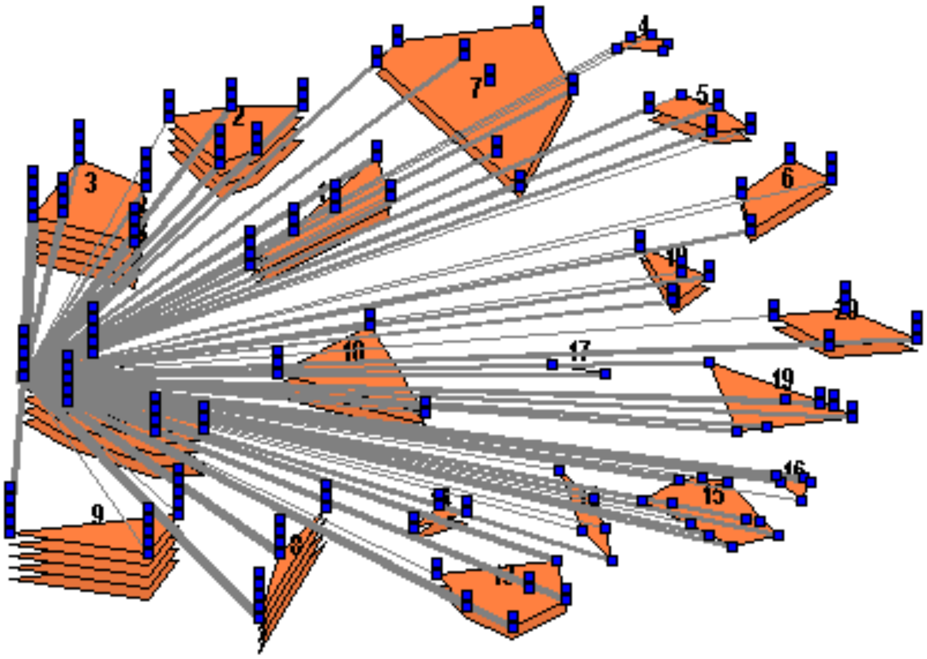

Figure 8: Spanning analysis for statement 61

Located in cluster 11, “guiding principles”, statement 61 “Evaluators must adhere to American Evaluation Association Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare” has the highest possible bridging value of 1.00. This value is so high partly because the statement has ties to all but 3-4 statements on the map. Also, there is a very strong tie from statement 61 to nearby, same-cluster statement 36: “Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy.” However, there are absolutely **no** ties from 61 to two of the other statements in its own cluster, namely statement 50 “Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight”, and statement 15 “Value findings of ‘no discernable effect’ as valuable feedback for program learning and redirection”. This suggests the cluster in which these statements sit is exceptionally lacking in conceptual cohesion, according to the responses of study participants.

Statement 61 has strong ties to statements in neighboring clusters, but also some fairly strong ones to statements in distant clusters. For example, in neighboring cluster 3, there a strong tie with 10 people sorting statement 61 to statement 2: “In an evaluation, the philosophical biases of the evaluator must be clearly identified”. Statement 61 also has strong ties to two statements in a neighboring cluster that relate to certification and rating of professional evaluators (statements 3 and 12). However, statement 61 also has moderately strong tie to the fairly distant statement 32: “There shall be a core set of evaluation policies which apply to all evaluations.”

To understand better the implications of these results for the conceptual coherence of the clusters and the overall usefulness of the map as a set of categories for a taxonomy, all of the statements on the map with bridging values of 0.75 or greater were examined for their connections to other statements as seen by sorters, and the strength of those ties. The goal was to see whether there was a common characteristic among these highly bridging statements that might explain their many connections to ideas in all areas of the concept map. In general, for evaluation policy ideas on which sorters did not agree, there is no single explanation for the disagreement. “Interloper” points appear to vary in the reasons for their cross-map connections, with some connecting in a principle-and-application way, some connecting in a type/sub-type way, and others simply connecting according to shared conceptual dimensions.

In cluster 3 “justice of evaluation process”, there are three evaluation policy ideas with high bridging values. The most striking example is statement 65 “The standard for the evaluation of all social programs shall be whether the program advances the goals of social justice”. This is a statement of principle with an exceptionally strong tie to statement 79 in cluster 1 “relevance of reporting”, namely: “Evaluations shall be conducted with the goal of assuring that the population or program benefits from the evaluation”. Here, statement 79 could be considered a *more specific sub-type or application of* statement 65.

In cluster 8 “certifying evaluator quality”, there are two statements with high bridging values. The first is statement 12 “Establish a "better business"-type consumer protection rating system for all evaluation companies, groups, and individuals working with publicly funded evaluations”. Statement 12 has moderate ties on the right side of map to low-bridging clusters populated with a mix of general and specific policies, namely cluster 12 “openness and democracy” and cluster 13 “roles and relations”. Statement 12 also has strong ties to cluster 16 “universal standards”, including one to statement 32, which expresses a central assumption of the study, namely: “There shall be a core set of evaluation policies which apply to all federal evaluations” (sorted together by 7 people). Statement 32 expresses a general idea asserting that there shall be standards for practice, while statement 12 expresses a more specific idea about *enforcing or making operational* such standards for practice. Statement 12 also has a strong tie to statement 20 “Standardize the evaluation language so that the evaluation work under one federal agency can be compared with the work under another agency” (in the view of 8 sorters) to statement 18 “The federal government shall establish a clear, universal working definition of the term "program evaluation" (sorted together by 7 people). Again, these two statements share a common theme of standardizing language of evaluation, with statement 18 a more specific *sub-type* of statement 20.

Statement 12 also has strong ties on the left side of map to cluster 11 “guiding principles”, including to statement 61 “Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare” (sorted together by 8 people) and to another statement about principles, statement 36 “Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy” (sorted together by 8 people). Again, one aspect of the general idea of high quality evaluator practice, enforced in statement 12, is the extent to which an evaluator adheres to accepted standards for the profession alluded to in statement 36.

In cluster 9, “capacity building”, statements with high bridging values include one more general statement: 40 “Build an evaluation culture”, and two somewhat more specific statements which could be considered *sub-types* of 40, namely: 97 “Provide technical assistance to foster evaluation capacity building within organizations and agencies”, and 96 “Support and nurture up-and-coming evaluators, through fellowships or sabbaticals at different agencies and opportunities for co-authorship.”

Statement 40 “Build an evaluation culture” has stronger ties on the lower right area of the map, including a strong tie to the general statement 32 “There shall be a core set of evaluation policies which apply to all federal evaluations” (sorted together by 7 people) in cluster 16 “standards”. Here, maintaining evaluation policies, as in statement 32, is one means of building of an evaluation culture, as in statement 40. Statement 40 also shows a similar ends-to-means tie to statement 23 “All federally funded programs must have a system in place for feedback and improvement” (sorted together by 8 people) in the distant cluster 15 “institutionalizing evaluation”. These connections could mean participants view evaluation policies and feedback systems as a means to achieve the end of an evaluation culture, or that evaluation culture and policies are considered part of an effective feedback system.

Statement 97 “Provide technical assistance to foster evaluation capacity building within organizations and agencies” shows a mix of weak and moderate ties to all clusters, however 97 has only a few moderate ties to closely neighboring clusters, as for example, to statement 14 "Evaluators shall serve as key members of the planning body for each project and program" (sorted together by 5 people). Statement 97 also has several moderate ties to statements in more distant clusters 14 "assuring the use of results", 13 "roles and relations", 12 "openness and democracy", and 15 "institutionalizing evaluation", and one very strong tie to statement 93 "Require training for federal, state and program managers--including what evaluation is, what constitutes effective evaluation work, and how to manage external evaluation" (sorted together by 11 people), sitting in distant cluster 13 "roles and relations". Since training in how to manage external evaluation is one specific aspect of evaluation capacity building in an organization, this could be considered a type/sub-type connection or an ends-to-means connection. Despite neatly fitting into the conceptual category of "capacity building", these ideas about evaluation technical assistance are clearly seen by sorters as having conceptual connections to many other types of evaluation policy in these results.

In cluster 11, "guiding principles", statements with high bridging values are statement 22 "Value evaluation partnerships between academia and low-resource communities" and statement 36 "Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy." While it is easy to see how the list of high principles in statement 36 might connect to many other policy ideas on this map, the high number of connections with statement 22, which calls for valuing a very specific type of evaluation partnership, is surprising, and may suggest participants simply weren't quite sure where to put it.

The last statement with a high bridging value was found in cluster 20 “intergovernmental collaboration”, statement 36 “Evaluators must adhere to the program evaluation standards as developed by the Joint Committee on Standards for Educational Evaluation (JCSEE) at the Evaluation Center, Western Michigan University, on utility, feasibility, propriety, and accuracy.” This statement has an exceptionally strong tie to same-cluster statement 61 “Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare”, with 16 people sorting 36 and 61 together. Since both statements reference established, general standards for practice well known among American Evaluation Association members, this strong tie between two statements of a roughly equal level of generality is neither type/sub-type or ends-to-means, but could be described as a peer connection.

In sharp contrast to the very strong peer-to-peer tie of 36 to 61 above, statement 36 (on JCSEE standards) has no ties to two other statements in its own cluster, according to sorters. These two other statements are statement 50 “Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight” and statement 15” Value evaluation findings of "no discernable effect" as constructive feedback for program redirection.” This means that although these statements appear in the same small cluster with statement 36, no participants sorted them together, calling into question the conceptual cohesion of this cluster.

However, statement 36 (JCSEE standards) has strong ties to immediately neighboring clusters, such as to statement 96 (in cluster 9 “capacity building”): “Support and nurture up-and-coming evaluators, through fellowships or sabbaticals at different agencies and opportunities for co-authorship" (sorted together by 7 people) and to statement 3 (in cluster 8 "certifying evaluator quality"): "Require that evaluators be certified in order to perform evaluations of publicly funded programs" (sorted together by 7 people) as well as to statement 12 "Establish a "better business"-type consumer protection rating system for all evaluation companies, groups, and individuals working with publicly funded evaluations" (sorted together by 8 people). All three statements, 96, 3 and 12 could be considered ways to operationalize or support implementation of the principles for practice found in the JCSEE standards mentioned in statement 36.

Statement 36 has a strong tie to statement 2 (cluster 3 "justice of evaluation process"): "In an evaluation, the philosophical biases of the evaluator must be clearly identified" (sorted together by 7 people), also to statement 73 (in cluster 1 "relevance of reporting"): "Evaluation findings shall explicitly address threats to validity"(sorted together by 6 people), as well as to statement 91 (in cluster 10 "integration of planning and evaluation"):"Every evaluation plan shall indicate how the results are to be used and communicated" (sorted together by 6 people). All three, statements 2, 73 and 91 could all be considered examples of standards for evidence or for good evaluation practice, and correspond to specific items under the general heading of the JCSEE standards.

Finally, statement 36 (JCSEE standards) has strong ties to a very distant statement expressing the central assumption of the study, namely statement 32 (in cluster 16 "institutionalizing evaluation"):"There shall be a core set of evaluation policies which apply to all federal evaluations" (sorted together by 7 people). Since 32 contains no specific topics or policy areas, it appears that the conceptual connection with JCSEE standards (statement 36) is again the idea of universal standards for evaluation practice.

Rating results

In concept mapping projects, participant rating inputs are sometimes used to eliminate lower-rated ideas from consideration, or shine a spotlight on highly-rated ideas for an action agenda. Here, the rating inputs are not used to eliminate statements or clusters, since the aim is to generate as comprehensive a taxonomy as possible. Instead, the rating inputs are gathered for their potential usefulness in identifying highly valued policies or policy categories to be addressed first in future applications of the taxonomy, either in organizational evaluation policy inventories or in research.

At the same time 400 AEA members were invited by the researcher to sort the 100 statements in this study, a separate random sample of 427 AEA members was also invited to rate the 100 statements on two scales, “merit”, and “feasibility”. Response to the rating invitation was stronger than for the more complex sorting task, with 66 invitees beginning the task for both of two rating scales, 63 completing for the first rating scale and 48 completing for the second rating scale. Participants’ rating responses on the “merit” and “feasibility” scales are shown in Figures 9 and 10 below, which depict more highly-rated ideas as taller stacks of points. Note that because the rating questions were asked using a 5-point, bi-modal scale, statements appearing with two or one dots in the stack were rated negatively or very negatively, on average. Statements with three dots reflect a neutral average rating. Statements with 4 or 5 dots in the stack received an average rating that was positive or very positive.

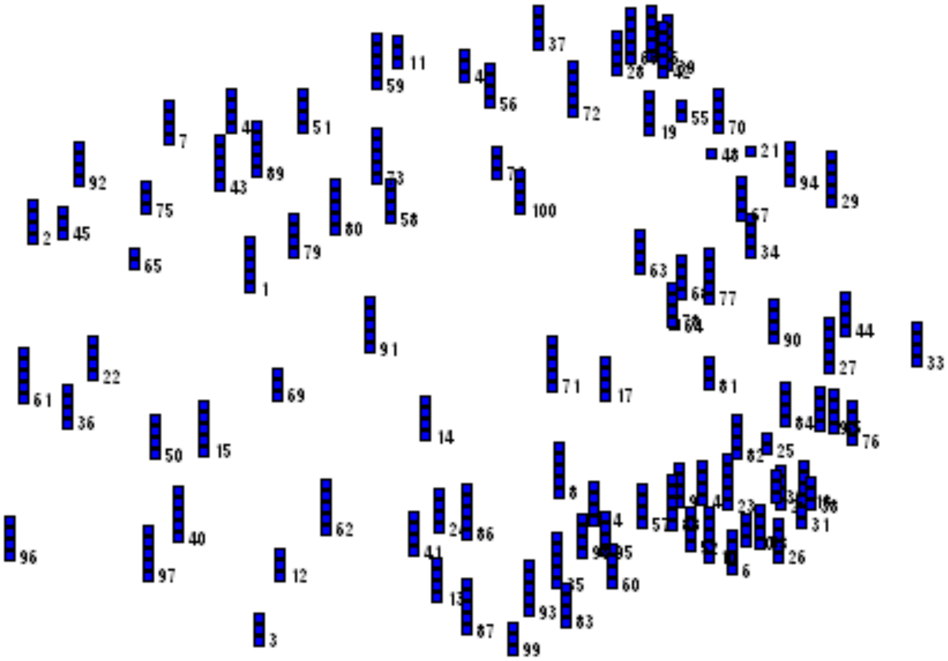

Figure 9: Point rating map for "merit" scale

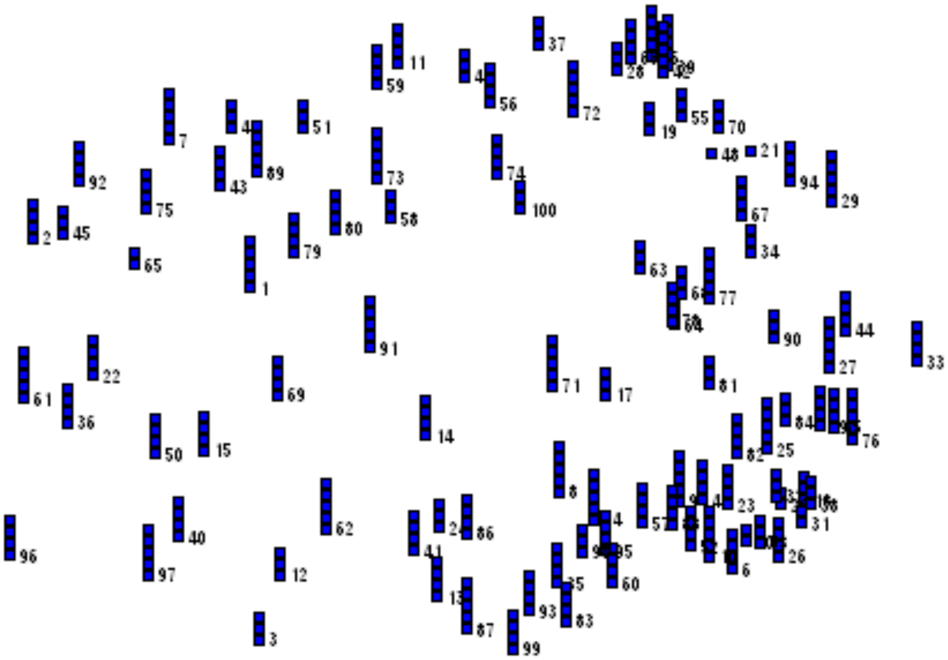

Figure 10: Point rating map for "feasibility" scale

The merit and feasibility point rating maps appear to be similar, suggesting that participant's ratings on these two different scales may be highly correlated. (Pattern match analyses discussed later in this chapter show this to be the case.) A preliminary, intuitive check of these results shows that a statement mandating experimental methods for all evaluations is rated low, while a statement calling for flexibility in choice of evaluation design is rated high. This matches expectations for a sample of American Evaluation Association members, as the organization is known for opposition to policies mandating the use of experimental design for all evaluations. Interestingly, the statement "There shall be **no** comprehensive set of evaluation policies" is rated low, while ratings for the statement "There shall be a set of comprehensive evaluation policies" is high. This suggests the group of participants who chose to rate statements tend to agree with one of the central assumptions of this study.

Next, average cluster ratings are examined to see whether any clusters were rated exceptionally low. To do this, the analysis combines rating inputs with the cluster solution. Figures 11 and 12 below show cluster rating maps for participant ratings of statements on the "merit" and "feasibility" scales, with clusters stacked higher to represent a higher average rating of statements within each cluster. On the cluster rating map for "merit", two clusters show exceptionally low average ratings. These are cluster 5, "strict standards for rigor" and cluster 8 "certifying evaluator quality". Four clusters received neutral average ratings of approximately 3 out of 5 on the Likert scale. These are cluster 3, "justice of evaluation process", cluster 16, "universal standards", cluster 18 "quality of evaluation practice", and cluster 19 "guidelines".

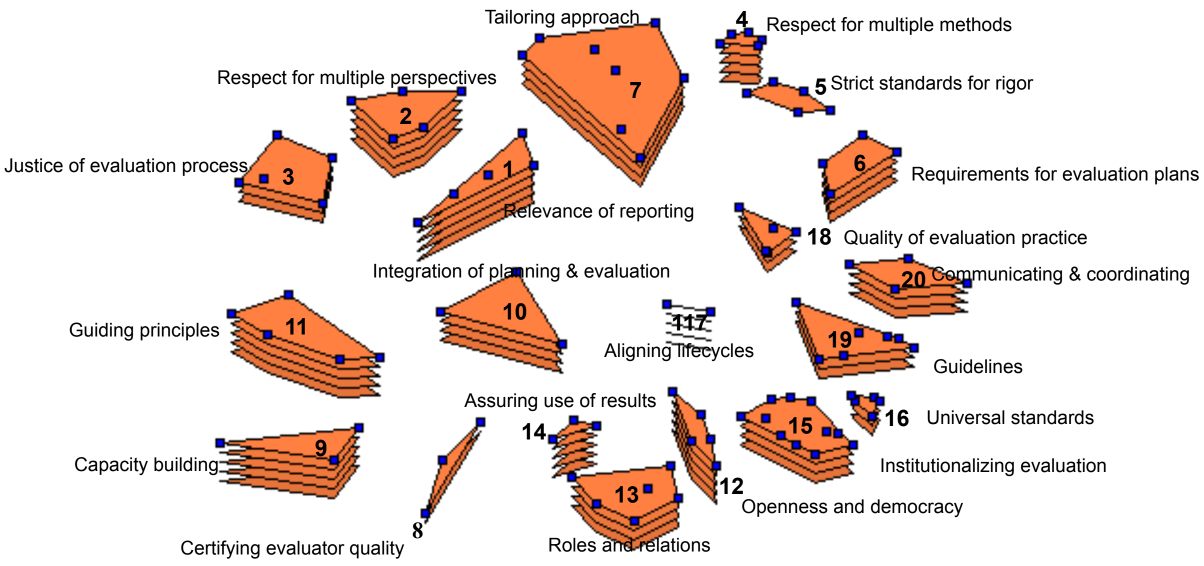

Figure 11: Cluster rating, "merit"

On the cluster rating map for "feasibility", cluster 5 "strict standards for rigor" and cluster 16 "universal standards" show exceptionally low average ratings. Cluster 3 "justice of evaluation process" receives a neutral average rating.

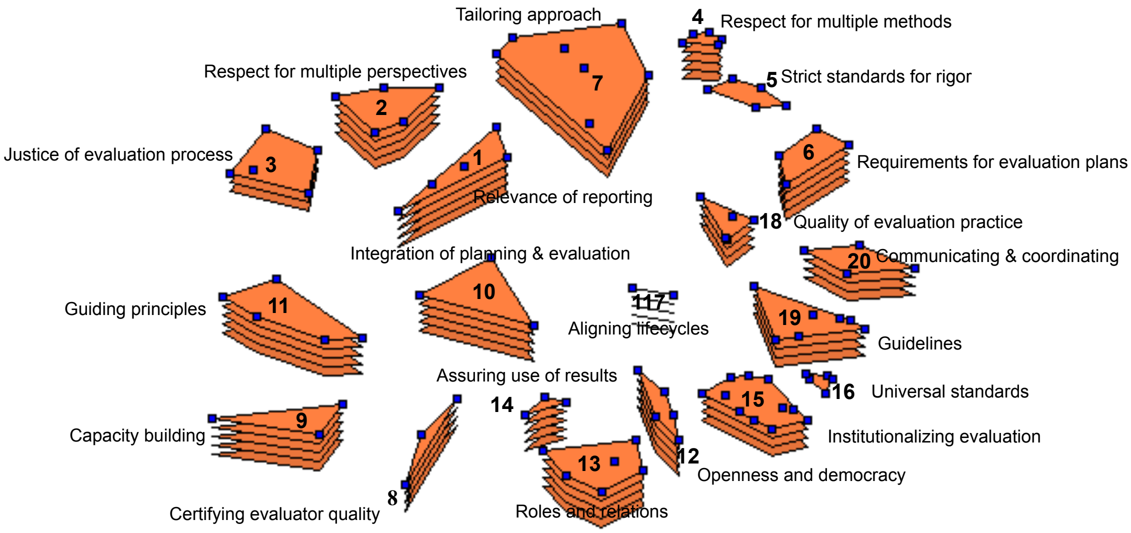

Figure 12: Cluster rating map for "feasibility" ratings

On these two maps, cluster 5 “strict standards for rigor” stands out as the lowest-rated cluster by far for both merit and feasibility. This cluster contains very specific methodological requirements for evaluations, such as “Evaluations shall be required to use a random experimental design”, which received the lowest average merit rating for any statement on the concept map, at 1.30 out of 5. This suggests these statements may be grouped together because participants viewed them as strikingly lacking in merit and/or feasibility.

Cluster 16, “universal standards” also has a strikingly low average feasibility rating compared to other clusters on the map. This low-bridging cluster contains statements such as “The federal government shall establish a clear, universal working definition of the term ‘program evaluation’” and “There shall be a core set of evaluation policies which apply to all federal evaluations”. Interestingly, this statement was rated just 2.65 out of 5 on feasibility.

Cluster 5 “justice of evaluation process” receives no better than a neutral average rating on both merit and feasibility. This cluster contains some strongly idealistic statements with which not all raters agreed, such as “The standard for the evaluation of all social programs shall be whether the program advances the goals of social justice.” (statement 65), and “Evaluation findings shall include a discussion of the minority composition of the group studied and an indication of whether the general findings apply to minority groups.” (statement 75).

Pattern matching analyses

In concept mapping, if a cluster's merit and feasibility ratings differ systematically, this raises questions about the nature of the statements in the cluster, why participants perceived this divergence, and what this could mean for the applicability of results. If merit or feasibility ratings differ systematically by rater characteristics, this invites questions about whether the map in fact represents a consensus, or whether further discussion among sub-groups would be needed to achieve a consensus result.

Statement ratings for this study are examined in more depth to probe for merit/feasibility divergences and for subgroup differences in ratings. A comparison of ratings on the two scales using a "pattern match" display confirms that "merit" ratings tend to correspond to "feasibility" ratings. A close look at ratings by participant characteristics finds almost no evidence that participants varied systematically by education, major work activity, main work setting, or country of primary residence in their ratings of evaluation policy ideas. The only exception is that raters who reported their primary work activity as "research" differed meaningfully from those who self-identified as "evaluators" in their relative average feasibility ratings of the different policy idea clusters.

Pattern matching is an approach to hypothesis testing in social science that looks not for a single effect A, but for an observed pattern of effects in relation to one another, for example a greater change in A than in B, and a greater change in effect B than in effect C. This approach is especially useful where an overall effect is not statistically significant due to lack of power, but meaningful changes in the interrelationships between two or more different expected effects may be observed. Rather than only looking at a change in absolute values, pattern matching looks at a group of effects relative to one another (Trochim 1985, Campbell, 1966). In concept mapping, a cluster pattern matching analysis “provides a comparison of average cluster ratings…[for] two separate stakeholder groups, [for] two different rating variables, such as impact and feasibility, …[or for] different points in time” (p. 19-20). A difference in ratings on merit versus on feasibility can suggest a difference between a participant group’s goals and their available resources to implement those goals. A difference in ratings among participant sub-groups can suggest a lack of consensus on goals (Kane and Trochim 2007).

Below is a pattern match analysis of merit rating responses compared with the feasibility rating responses for all raters (Figure 13).

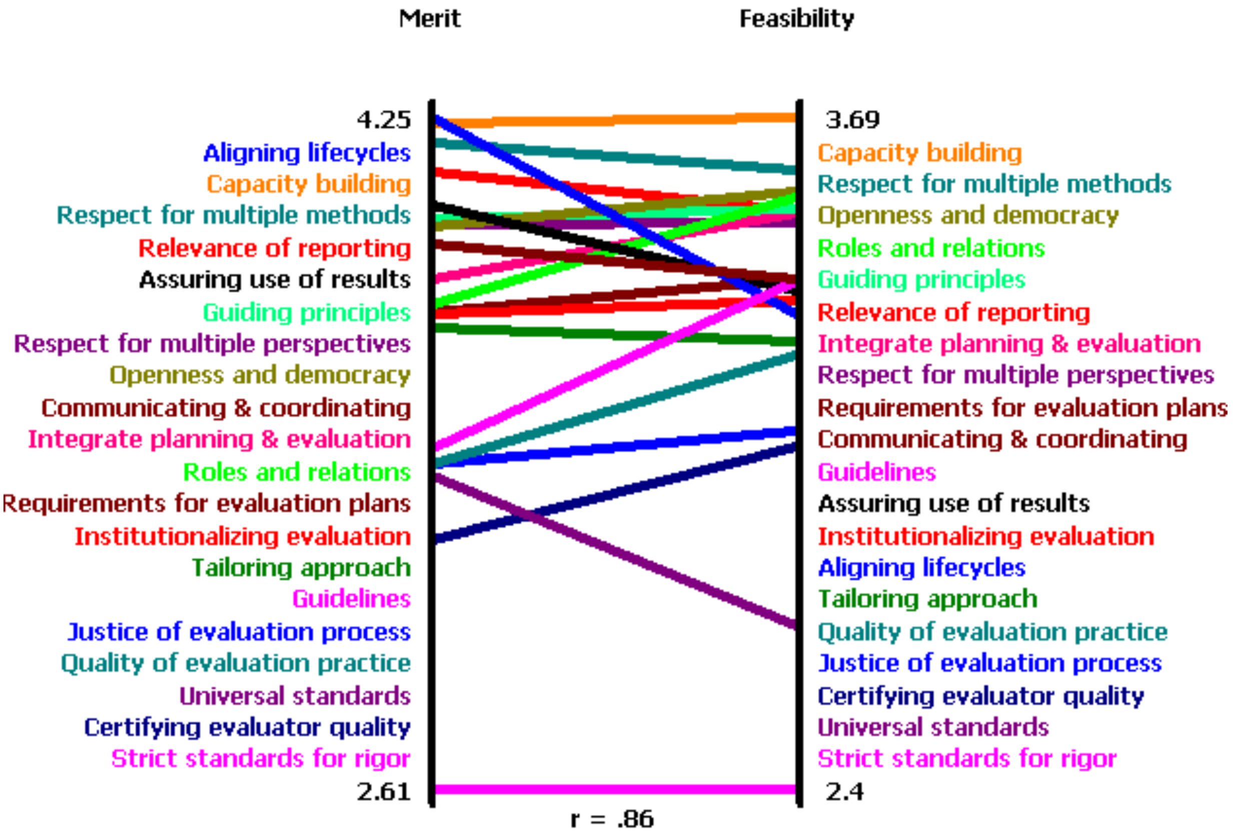

Figure 13: Merit versus feasibility, cluster average ratings

In this pattern match, which is a comparison of merit and feasibility ratings by cluster, a few clusters show ladder “rungs” with a slope greater than zero. For example the “aligning to program lifecycle” cluster and the “universal standards” cluster, both rated relatively higher on merit than on feasibility, and the “guidelines” cluster, rated higher on feasibility than on merit. This might suggest that although participants feel there should be some universal standards for evaluation, they see as more feasible guidelines that would empower lower levels of the government to set specific evaluation policies. The downward slope of the “aligning to program lifecycle” rung suggest that while participants agree evaluation method should match program maturity level, they see attempts to make this a policy as infeasible. However, overall, most of the rungs of the ladder display are close to parallel, and the r value is high (at .86), which suggests that where the average cluster rating on “merit” is higher (relative to merit rankings for other clusters), that cluster’s average cluster ratings on “feasibility” tend to be higher (relative to feasibility rankings for other clusters). Where the average merit ratings for a cluster are relatively lower, its average feasibility ratings tend to be relatively lower as well.

Since the American Evaluation Association has a higher proportion of evaluation practitioners than evaluation researchers, and other organizations with more evaluation researchers were not included in the samples for this study, it seemed interesting to look at differences among sub-groups by primary work activity was “research” and those for whom it was “evaluation”.

In the pattern match graphs below, differences in rating inputs are examined for each rating scale, comparing inputs from participants who identify their main work activity as “research” (n=4) with inputs from participants who identify their main work activity as “evaluation” (n=27).

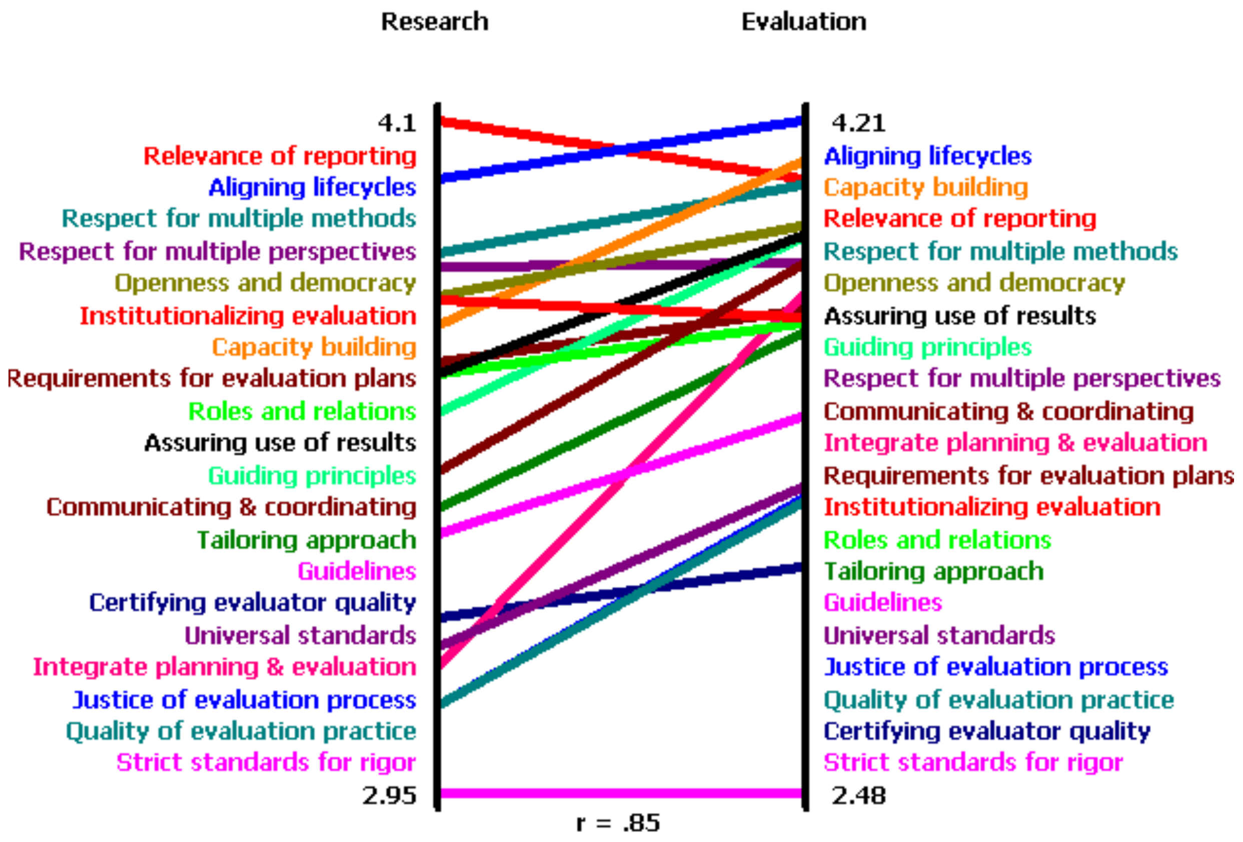

Figure 14: Research versus evaluation, cluster average “merit” ratings

Comparing merit ratings between researchers and evaluators, the pattern match display shows some rungs sloping upward, most notably the “integrate planning and evaluation” and “quality of evaluation practice” clusters, suggesting evaluators rated these clusters more highly on merit than did researchers. However, the high r value of .85 indicates that overall, the two rating patterns are highly correlated, meaning the two groups ranked clusters roughly the same on merit.

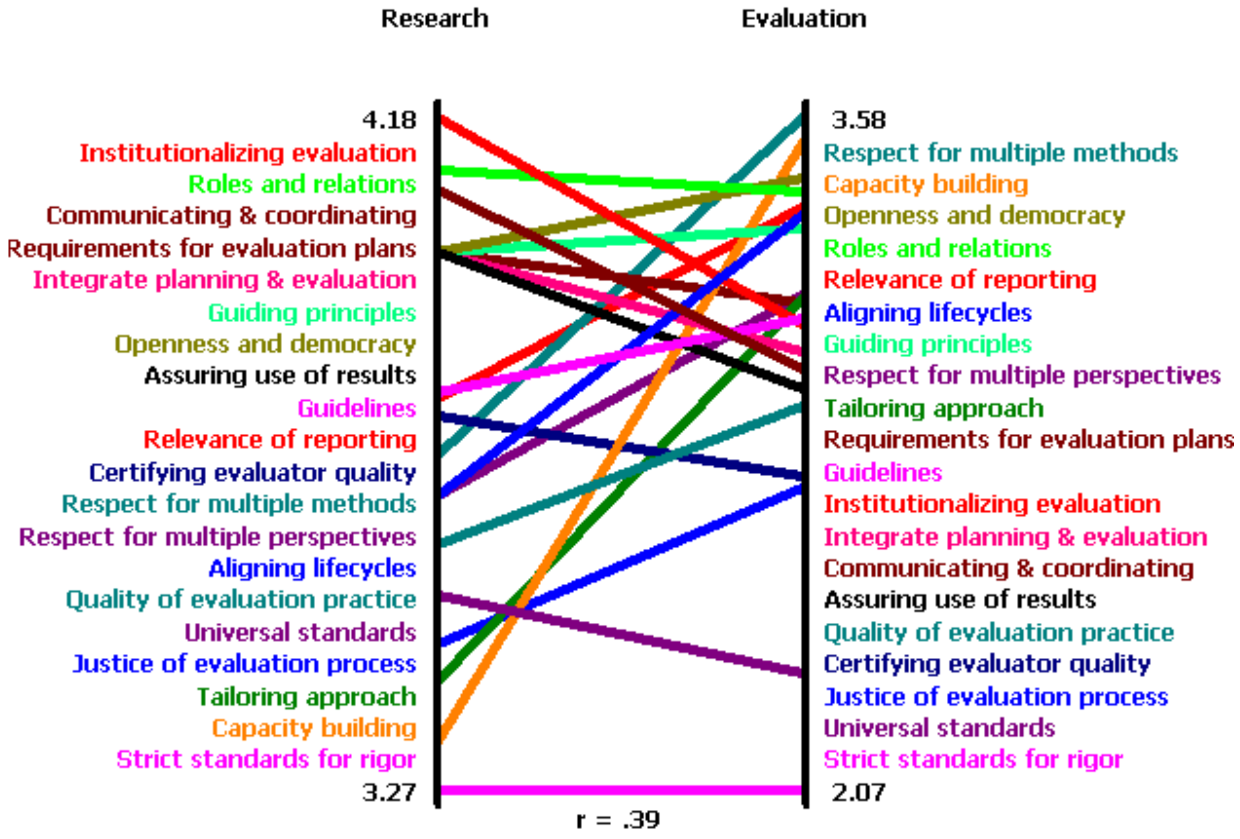

Figure 15: Research versus evaluation, cluster average “feasibility” ratings

On the other hand, differences in ratings on the “feasibility” scale appear meaningfully different between the two groups, with an r value of .39. This pattern match shows some potentially meaningful differences in the relative average cluster ratings on feasibility by these two sub-groups for most clusters. A striking example is the “capacity building” cluster, rated relatively quite low on feasibility by researchers but relatively quite high on feasibility by evaluators, perhaps reflecting evaluators’ higher level of optimism that non-specialists can learn to do good evaluation. Likewise the “tailoring approach” cluster is rated quite low on feasibility by researchers relative to their ratings of other clusters, but quite high on feasibility by evaluators relative to their ratings of other clusters, perhaps reflecting a greater leaning toward methodological flexibility among this sub-group. The lowest rated cluster for feasibility by both groups is the “strict standards for rigor” cluster, suggesting some consensus across researchers and evaluators in this sample on the need for some flexibility in methods. The highest rated cluster for evaluators was “respect for multiple methods”, possibly reflecting the ongoing debate about requiring experimental methods. Researchers rated the “institutionalizing evaluation” cluster, with its statements about high-level oversight of internal evaluation, most highly, perhaps reflecting their view that experts need to be in charge of evaluation.

Recall that many original brainstorm participants responded saying they were unsure if they should participate since they were not from the United States. When study invitees wrote to the researcher about this, they were encouraged to participate in the study nonetheless. While many who had expressed this hesitancy did not end up participating, some did. To examine possible differences in ratings by those based in the U.S. as compared with those not based in the U.S., a pattern match analysis was conducted to look at relative ratings between the two groups (primary residence U.S.A., n=52 ; primary residence not U.S.A. n=8), shown below.

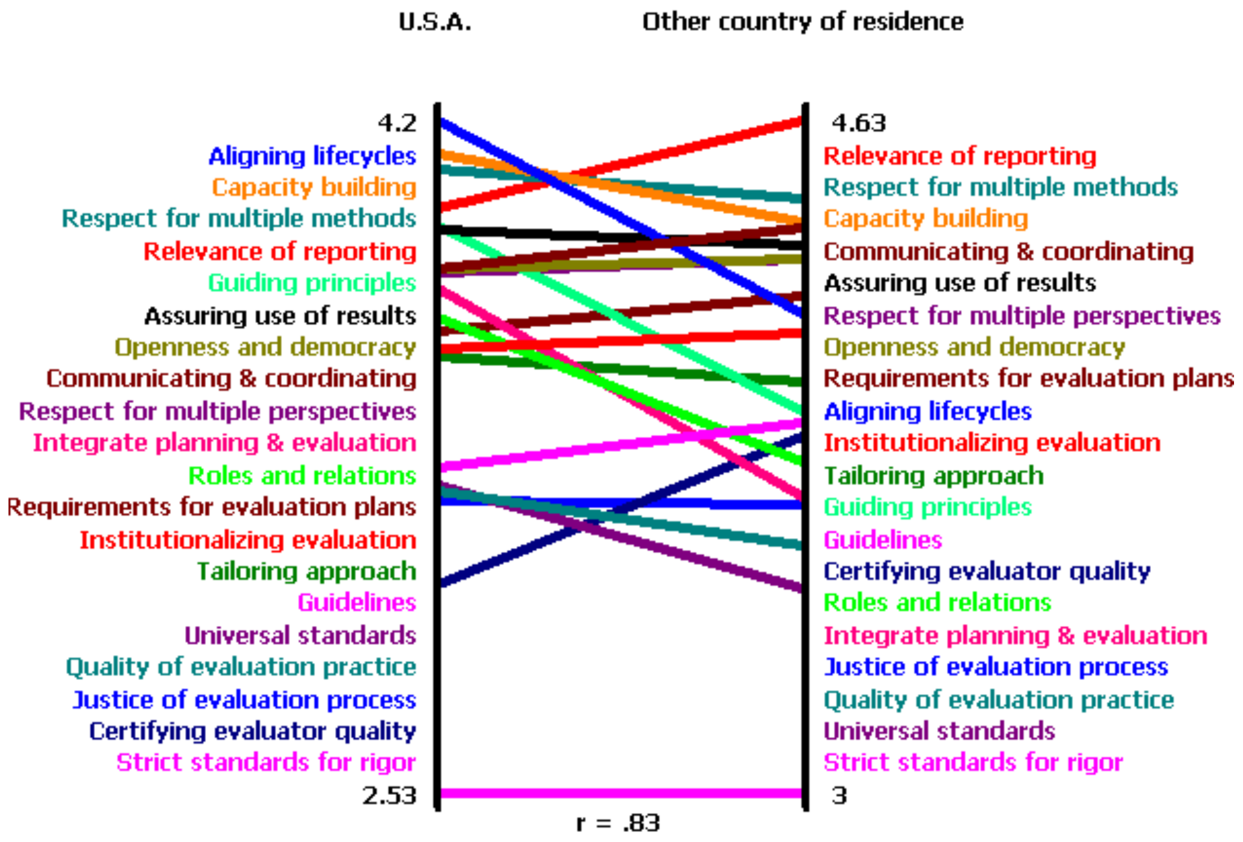

Figure 16: U.S. residents versus non-U.S. residents, cluster average "merit" ratings

U.S. participants' cluster ratings for "merit" appear different from those of non-U.S. participants, with some clusters rated relatively higher and others rated relatively lower by the two groups. However, the high r value of .83 suggests that the two sets of ratings are not meaningfully different.

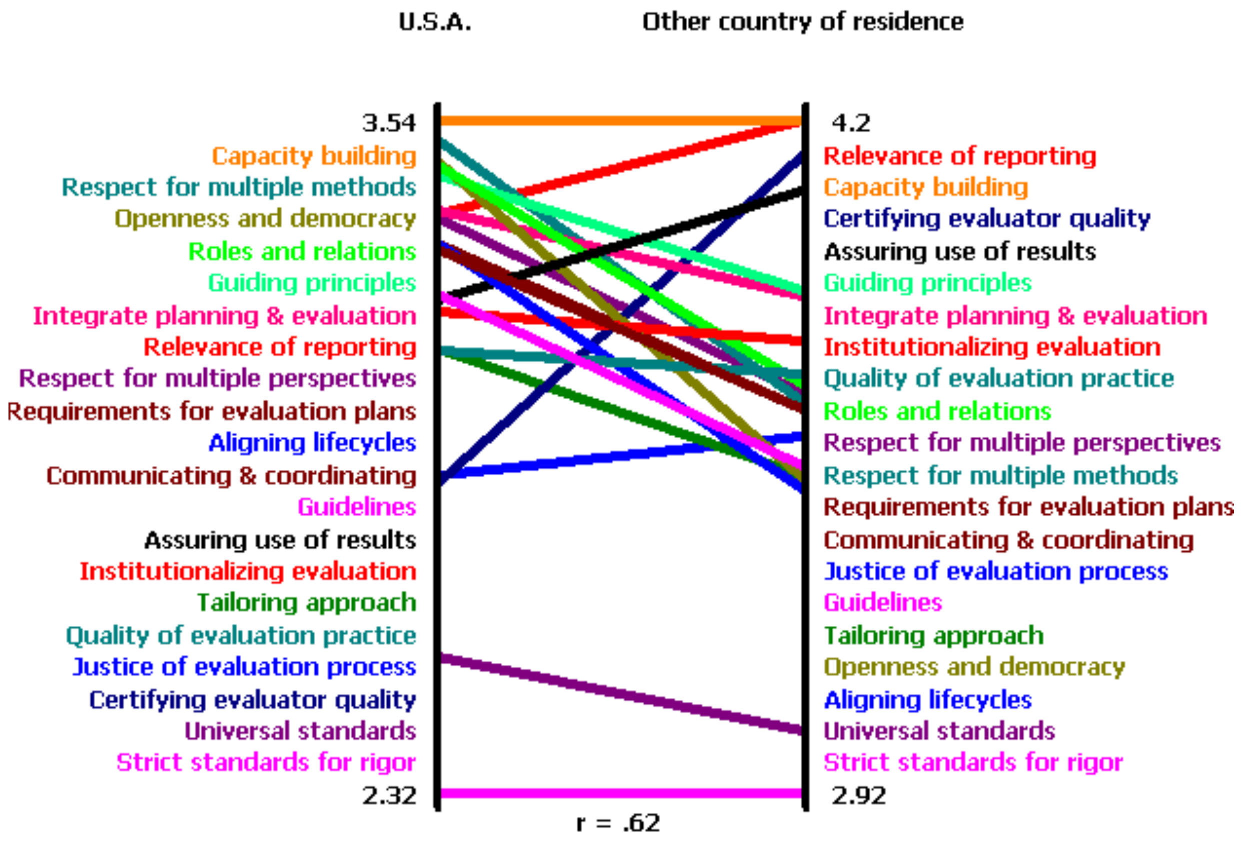

Figure 17: U.S. versus non-U.S., cluster average "feasibility" ratings

A comparison of rating patterns for these two groups on "feasibility" reveals a somewhat lower r value of .62. The strongest difference in relative ratings is for the "certifying evaluator quality" cluster, with non-U.S. residents rating this cluster much higher, in relative terms, than U.S. residents. This may reflect the fact that other countries have clearer standards for professional certification of evaluators than does the U.S.A. The r value, while lower than for merit for these two sub-groups, is still relatively high, suggesting a strong correspondence of ratings by resident and non-resident participants.

Next, a “go-zone” analysis was conducted to see which evaluation policy ideas were considered both highly important **and** highly feasible by study participants.

Go-zone analyses

Identifying statements rated highly on both merit and feasibility can help streamline and prioritize evaluation policy inventory activities by focusing attention on policy areas considered most “actionable” by stakeholders. In the “go-zone” display below in Figure 18, statements rated above average on both merit and appear in the upper right quadrant of the graph.

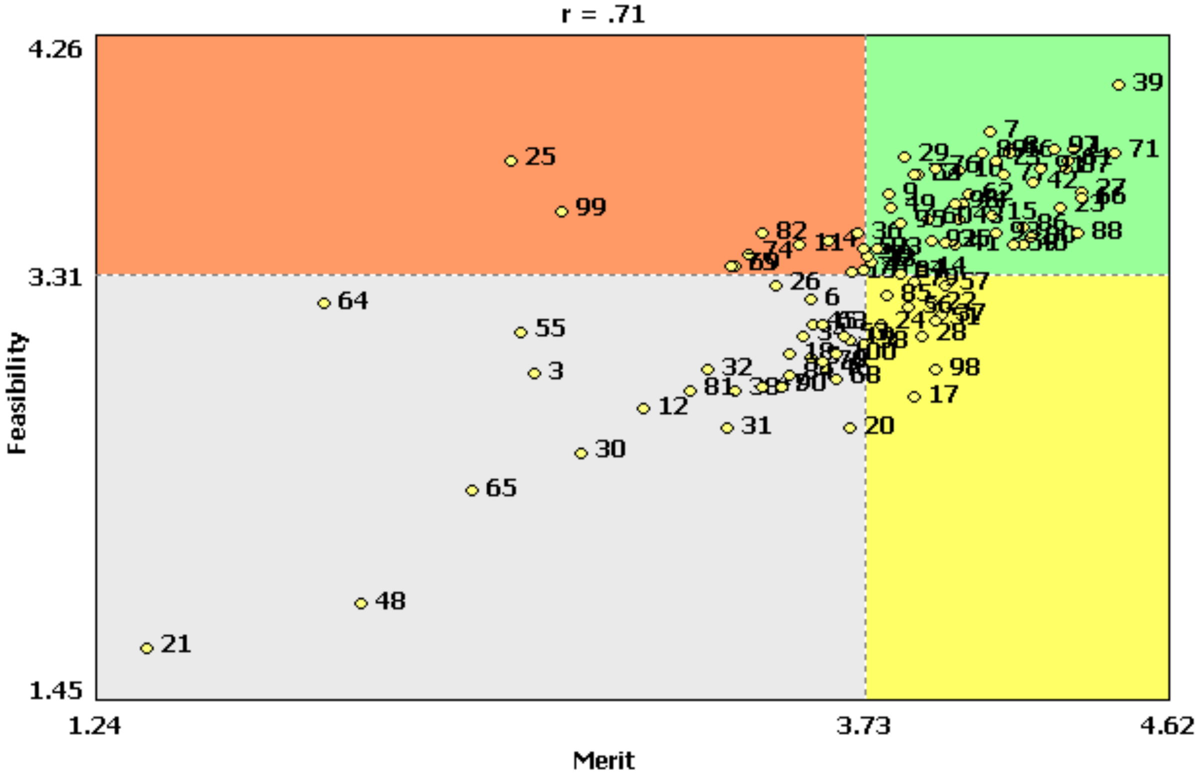

Figure 18: Go-zone, statement ratings for “merit” versus for “feasibility”

The r value for this analysis is .71, reflecting the fact that in this study, across all statements, merit ratings are fairly highly correlated with feasibility ratings. The only two clusters not represented in this overall “go-zone” are cluster 5 “strict standards for rigor” and cluster 16 “universal standards”, reflecting, perhaps, a hesitation by study participants about policies that may be too rigid or too broadly applied. Forty seven statements appear in quadrant one of this go-zone analysis, which might be too many to consider at once.

To narrow the group of ideas for easier examination, for each statement in the “go-zone”, merit and feasibility rating scores were added to together, and the top-rated twenty statements were identified. A list of these statements by cluster appears in Appendix F. In general, these statements lean toward more flexible guidelines, leaving out the more rigid rules from the 100 statements offered for sorting. Statement 32, which calls for a comprehensive set of evaluation policies, is notably missing from this short list of the very most actionable evaluation policy ideas.

Clusters not represented in the top twenty rated statements include cluster 3 “justice of evaluation process” (though many of these ideas are captured in cluster 11, “guiding principles”), cluster 5 “strict standards for rigor”, cluster 7 “tailoring approach” (though similar ideas are included from the “respect for multiple methods” cluster), and cluster 16 “universal standards”. Strict or universal standards are clearly out of favor with these participants, who may be concerned with preserving evaluator prerogatives.

CHAPTER 5: CONCLUSIONS

The main purpose of this study has been to generate a grounded taxonomy of the construct of evaluation policy, using survey responses from a large sample of evaluation professionals. Intended applications of this framework include the inventory of evaluation policy in organizational contexts, and the construction and testing of theories of evaluation policy impact.

Related to that purpose, this study addresses two questions. First, what is the definition of evaluation policy? What are all the relevant types of evaluation policy in the U.S. federal government context? Second, how do these results compare with previous framings of this construct, specifically the intuitive definition of evaluation policy and taxonomy of evaluation policy types of Trochim (2009) and the content of the AEA Roadmap?

This study makes the following contributions to previous theoretical taxonomies of evaluation policy: 1) First, it generates, from the responses of evaluators and evaluation researchers, a fuller and clearer definition of evaluation policy; 2) Second, it adds a new “values” dimension and a new sub-type of “coordination” to the Trochim (2009) evaluation policy taxonomy; 3) Third, it generates classification criteria for each of the categories in the evaluation policy taxonomy; 4) Last, it proposes an alternate taxonomy incorporating the policy dimensions guiding value and mechanism, with practical, step by-step guidance for its use by evaluators working with organizations seeking to develop their evaluation policies.

Definition of evaluation policy

The over 920 evaluation policy ideas offered by study participants include the full range of evaluation types, including needs assessment, evaluability assessment, implementation/process evaluation, fidelity assessment, benefit/cost analysis, outcome evaluation, impact evaluation. These data also contain evaluation policy ideas referring to all stages of evaluation, including planning, data collection, analysis, reporting and utilization. Ideas in these data refer not just to the processes involved in "doing evaluation", but also to the processes of capacity building and setting up institutional structures to support evaluation. In contrast to the idea that policies are limited to "principles" or "rules", the responses to this survey contained numerous policy ideas setting up standards or guidelines, none of which contained language about sanctions for non-compliance. Many of the evaluation policy ideas offered in the brainstorming phase of this study use language of "standards" or "guidelines. One or two mention positive incentives for desired evaluation practices.

Recall that the Trochim (2009) definition of evaluation policy was: *"Evaluation policy is any rule or principle that a group or organization uses to guide its decisions and actions when doing evaluation."* Based on the results in this study, two refinements to this definition are here proposed. First, the definition is extended to included forms beyond "rules or principles" to include *standards or guideline*s to capture an important layer of flexible policies in between principles and rules, for which more specific details are meant to be delegated to a lower level on the organizational hierarchy. Second the object of evaluation policy is extended beyond simply "decisions and actions when doing evaluation" to include "decisions and actions when planning, conducting, reporting or using evaluation, or which establishes institutional processes or structures to maintain an evaluation system”. The new definition reads: *“Evaluation policy is any principle, rule or standard used to guide an organization’s decisions and actions in planning, conducting, reporting or using evaluation, or any policy which establishes the organizational capacities, processes or structures for an evaluation system.”*

Comparing taxonomies

The concept map in this study shows a marked correspondence to the categories in the Trochim (2009) taxonomy and the AEA Roadmap (2009/2010). This means that when aggregated, the twenty sets of sorting inputs in this study correspond fairly closely to those of the intuitive taxonomy of Trochim (2009), as well as to the main topic areas in the AEA “Roadmap” document. Only two clusters, cluster 11 “guiding values” and cluster 3 “justice of evaluation process”, could not be easily matched to types in this intuitive scheme. For a three-way comparison of taxonomies, please see Table 5 below.

Table 5: Comparing taxonomies

Trochim (11/09) policy wheel	**AEA Roadmap (2/09 and 10/10)**	**Concept mapping clusters 11/09**
1. Goals	**PURPOSES OF EVALUATION** use program evaluation to… (p. 4)	
2. Participation	**QUALITY** *develop quality standards*	8 certifying evaluator quality
who does evaluation	professional competence	12 openness and democracy
		2 respect for multiple perspectives
who makes policy	who makes policy? = agency evaluation coordinators?	15 institutionalizing evaluation
3. Capacity	**MANAGEMENT** *assign senior officials, prepare evaluation plans, provide sufficient funding, ensure support for evaluation units & staff*	9 capacity building
4. Management	**SCOPE & COVERAGE**	17 aligning lifecycles
staff time and financial resources	resources; scope; coverage	14 assuring use of results
5. Roles responsibilities of different people for evaluation	**MANAGEMENT** **INSTITUTIONALIZING EVALUATION**	13 roles and relations
	Executive branch role: Congress' role; cross-branch collaboration	16 universal standards?
	INDEPENDENCE	
	integrate evaluation into program management	10 integration of planning and evaluation
6. Process and Methods question identification, sampling, measurement, design, analysis	**QUALITY** analytic approaches and methods	4 respect for multiple methods 5 strict standards for rigor 8 quality of evaluation practice
		7 tailoring approach
	MANAGEMENT evaluation plans	6 requirements for evaluation plans
	MANAGEMENT evaluation policy & procedures	19 guidelines
7. Utilization	**TRANSPARENCY** disseimination of results	1 relevance of reporting
8. Meta policy	**MANAGEMENT** *agency evaluation coordinators or centers*	14 assuring use of results
evaluation of evaluation	**MANAGEMENT** Coordinate and communicate about evaluation efforts across agencies	15 (institutionalizing evaluation)
		20 communicating and coordinating
	QUALITY	11 guiding principles
	AEA guiding principles	3 justice of evaluation process

Based on these results and the correspondences in the grid above, two enhancements to the Trochim (2009) taxonomy are proposed. First, “values” is added as a third dimension, alongside evaluation system component and hierarchical level in the Trochim (2009) scheme. The theme of “values” emerges strongly in these results, both in cluster 3 “justice of evaluation process”, with statements about upholding the goal of social justice in how evaluation is conducted, and in cluster 11 “guiding values”, with statements about values in evaluation, and how values guide the way evaluation is to be done and results used. Cluster 11 includes statements that reference both the AEA Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare, and the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy. Other statements address the need to bring out or attend to minority views or the impact of programs on minority groups. However, values are not isolated in these two clusters. The spanning analyses in this study suggest many other evaluation policies are related to values. The interpretation is that every evaluation policy is informed by some guiding value, such as accountability, transparency, equity, efficiency, justice, or respect for culturally different ways of understanding the world. So all evaluation policies could be usefully cross-indexed by their underlying values.

Second, “coordination” is added as a main sub-type of the Trochim taxonomy category of “meta-policy”, alongside “evaluation of an organization’s evaluation activities”. The “Roadmap” places at the agency level the responsibility for coordinating across agency boundaries, suggesting that agency evaluation coordinators or evaluation centers be in charge of assuring quality in evaluation. In contrast, the “Roadmap” includes no mention of a central coordinating body reporting to the Executive and Congress. Instead, it recommends that agency evaluation coordinators “consult with Congress” on defining policy and program objectives (AEA Roadmap, p.5). The “Roadmap” includes calls for agencies to “coordinate and communicate about evaluation efforts across agencies with overlapping or complementary missions” and “develop written evaluation policies across and within federal agencies that can guide evaluation efforts and help ensure their quality”. Here, the “Roadmap” appears to be urging federal agencies to voluntarily coordinate with other agencies. Interestingly, this document does not call for its own content, which reads like a national evaluation policy, to be formalized as an over-arching policy.

In contrast to the absence of central coordination in the “Roadmap”, some of the brainstormed evaluation policy ideas in these results call for a more centralized, coordinated oversight of evaluation activities. The initial set of survey responses includes three statements giving such oversight functions to a person or body. Each suggests a different institutional structure for who will be responsible for evaluation. The first would set up “an independent agency” to monitor attainment of goals of other agencies. The second would establish “a ‘statutory office’ in the Office of Management and Budget that is focused on evaluation policy”. The third states “A ‘chief program evaluation officer’ (CPEO) shall be appointed for each federal agency, and shall serve on a council with reporting requirements to Congress and the President”. (Note that this provision appears in a U.S. Executive Order from 2007 for performance measurement functions).

Survey responses directly address coordination in other ways as well. First, four statements in these results say evaluation should take place on multiple levels of a program system, including local, state and federal, as well as evaluation and meta-evaluation. One statement stands out as especially systems-oriented: “Value the levels of evaluation - at least local, state and federal - how they are linked, how they differ by function, and how they are all needed to contribute to a comprehensive evaluation of complex systems change.” Second, three statements call for coordination of evaluation policy among local, state and federal levels, echoing the Trochim “inheritance of policies” principle, which calls for a seamless hand-off by more general policies at higher levels to more specific policies at lower levels. Third, several of the statements in these results call for central policy on coordination of **across** peer sub-units of an organization. These include: “There shall be a core set of evaluation policies which apply to all federal evaluations”, “Standardize the evaluation language so that evaluation work under one federal agency may be compared with the work under another agency”, and “Where appropriate, agencies and programs shall use identical measures across programs and agencies.” One interpretation of this is that evaluators are likely to be concerned not just with establishing policies that encourage evaluation and its use at the agency level, but also with the consistency and coordination of evaluation policies across agencies, due to 1) the increased public utility of such coordination and 2) the greater ease in professional interactions for evaluators working across different federal settings. To incorporate the idea of standardization across an entire complex organization, as expressed in these data, a new sub-type of “coordination” is added to the “meta-policy” category of the Trochim (2009) taxonomy. Here, “coordination” refers not only to the harmonious inheritance of policies through successive layers of a hierarchy, but also to **horizontal** coordination, in the sense of concerted action across an organization’s peer sub-units, such as federal agencies.

It should be noted that, in contrast to ideas about more central control or standardization of evaluation, the original set of survey responses for this study contains a small but articulate group of voices expressing direct opposition to the idea of evaluation policy and to having a "comprehensive set" of evaluation policies. Two statements in the original set of responses oppose having a "comprehensive set" of evaluation policies. Three statements say it is impossible for any policy to capture or fit all situations or contexts. One statement talks of the practical obstacle of evaluation policies being un-enforceable. Another suggests that in lieu of having evaluation policies to direct decisions about evaluation, decisions be made case-by-case, through a consensus of stakeholders. One says only that the government should reduce the number of policies, since they are associated with excess "paperwork" burdens. A decision not to have a coordinated system of evaluation policies or evaluation policies at all is, itself, a policy. Because it represents a decision not to coordinate, even such an "un-policy" arguably belongs under the meta-policy sub-type of "coordination".

Studies of evaluation systems in literature reviews for this study largely confirm the Trochim (2009) taxonomy as well as the importance of horizontal coordination. The table below shows how the eight categories in the Trochim (2009) evaluation policy taxonomy map to key elements identified in theoretical and empirical studies from the literature review for this study. A new sub-type of "coordination" is added to the "meta policy" category, with corresponding dimensions from the literature shown in the table below.

Table 6: Trochim (2009) categories
By evaluation system characteristics from the literature

category	**evaluation system characteristics**
Goals	evaluation demand (2)* articulation of objectives organizational uses of evaluation purpose of evaluation results orientation
Participation	stakeholder involvement
Capacity building	organizational culture evaluation supply internal capacity IT resources financial and political support staff knowledge, skill and attitudes (2) evaluation networks evaluation training
Management	evaluation planning
Roles	roles (4)
Process and methods	procedures epistemological perspective; planning (2) study design
Utilization	evaluation demand use (5)
Meta-policy	inventory of studies
Coordination	*institutional integration* *intra-organizational coordination* *connecting supply and demand*

*Number in parenthesis after item indicates how many studies included this item.

Many of the components of evaluation systems in the literature correspond to the categories in the eight-part evaluation policy taxonomy of Trochim (2009). The "goals" category corresponds to demand for evaluation as in Liverani and Lundgren and Mackay (2007), the articulation of program objectives as in Wholey (1970), the four organizational uses of evaluation articulated by Shuamberg-Muller (2005), the purpose of evaluation, as in Debelstein and Rebien (2002) and "results-oriented" goals as in Mackay, 2007 and Furubu, Rist and Sandahl, 2002. The “participation category” corresponds to the diversity of evaluation suppliers highlighted by Summa and Toulemond (2002) and by Furubu, Rist and Sandahl (2002), as well as the involvement of stakeholders as described in Summa and Toulemond (2002). The “capacity” category corresponds to organizational culture, evaluation supply, internal capacity, information technology resources, financial and political support, staff knowledge, skill and attitudes, evaluation networks and evaluation training mentioned in Liverani and Lundgren (2007), Leuuw (2008), Dahler-Larsen, 2006, Chelimsky, (2009) and Summa and Toulemond (2002). The “management” category corresponds to evaluation planning, as in Wholey (1970), Debelstein and Rebien (2002) and Summa and Toulemonde (2002). The “roles” category is highlighted by Debelstein and Rebien (2002) and by Summa and Toulemond (2002), and corresponds to evaluation suppliers and demanders, as in Liverani and Lundgren (2007) and Furubu, Rist and Sandahl (2002). The “process and methods” category corresponds to procedures as described in Liverani and Lundgren (2007), epistemological perspective (see Leuuw 2008) and planning and study design as in Debelstein and Rebien (2002). The “utilization” category is mentioned in many studies, including Liverani and Lundgren (2007), Leuuw (2008), Wholey (1970), Shuamberg-Muller (2005), Debelstein and Rebien (2002) and Summa and Toulemond (2002). “Meta-policy” corresponds to the inventory of studies mentioned in Summa and Toulemond (2002).

Strikingly, the idea of hierarchical or “vertical” integration of evaluation policies through different levels in an organization’s hierarchy is barely mentioned in these studies. In the few places where the relationship of higher-level policies to lower-level policies is considered, it is in empirical studies designed to see whether higher-level policies have any effect on evaluation policies and practices at lower levels, as in evaluations of European Union and OECD DAC policies and policy guidance, as to their effect on member country policies. This may have more to do with the level of analysis of the studies selected than anything else.

Harmonization between the evaluation unit and other institutional functions, or among parallel organizational units within an organizational system, is mentioned often in these studies as a key element in building a healthy and successful evaluation system. Specifically, the literature mentions the importance of: 1) institutional integration, defined as how the evaluation function is situated within the organizational hierarchy; 2) horizontal coordination, defined as structures for networking or collaboration among parallel sub-units of the organization and 3) feedback response structures, defined as arrangements to connect evaluation supply with evaluation demand. Designation of a new sub-type of “coordination” under the Trochim “meta policy” category would assist evaluation system designers by encouraging them to consider these crucial relational elements, and to consciously develop clear policy to guide action in this area.

Examining evaluation systems and policies from other contexts (whether national, international or supra-national) can inform a U.S. taxonomy of evaluation policy. However, the types of evaluation policy needed may depend, at least in part, on a U.S.-specific definition of evaluation, and on the problems and opportunities specific to the organizational environments in this country. (Furubu, Rist and Sandahl, 2002)

Operationalizing the taxonomy

The cluster map can be seen as largely confirming the Trochim (2009) taxonomy. In this interpretation, the statements in those clusters provide specific examples to help refine the sensitizing concept of evaluation policy and operationalize the intuitive set of policy types of Trochim (2009). For example the Trochim category "participation" corresponds to the concept mapping clusters "openness and democracy", "certifying evaluator quality", "respect for multiple perspectives", and certain statements in the cluster called "institutionalizing evaluation". The Trochim category "process and methods" corresponds to the concept mapping clusters "respect for multiple methods" and "strict standards for rigor", "requirements for evaluation plans", "tailoring approach", "requirements for evaluation plans", and certain statements in the clusters "universal standards", "guidelines" and "quality of evaluation practice". Trochim's "management" category corresponds to concept mapping clusters "aligning lifecycles", and certain statements in the "universal standards" cluster. Statements each of the concept mapping clusters could be used as a foundation on which to build inclusion rules for each category.

What follows is a proposed categorization scheme in the form of a set of simple inclusion rules, based on the Trochim 8-part taxonomy with a new "values" dimension and "coordination" sub-type added. For the most part, only those statements included in the set of 100 that was sorted by participants are included. However, in a few cases, where the synthesis and reduction process cut out unique statements useful in elaborating a Trochim category, policy ideas from the original statement set for this study not included in the reduced set for the concept mapping analysis are included here. These are indicated in italics.

1. Goals: an evaluation policy that belongs in this category is a direct expression or assertion of the organization's view of the function of evaluation within the organization. *Example goals include: accountability, oversight, program management, quality improvement, providing useful information to the primary intended users, informing decision making by policy makers, including funding and program termination decisions and tracking the long-term impact of programs.*

2. Participation: An evaluation policy that belongs in this category pertains to who is qualified to do evaluation in the organization, or whose perspectives are considered in designing evaluations, or who is consulted in evaluation funding decisions, or who oversees evaluation activities.

3. Capacity: An evaluation policy that belong in this category relates to training of an organization's staff in evaluation practices or technical assistance to staff in performing evaluation functions, or efforts to build an evaluation culture within an organization.

4. Management: An evaluation policy that belongs in this category addresses the *funding of evaluation. Examples include set percentages for evaluation in project budgets*; policies on the strategic use of evaluation resources as appropriate to the lifecycle of the program.

5. Roles: An evaluation policy that belongs in this category establishes the role evaluators or for those conducting management or oversight of evaluation activities, or intended interactions among any of these roles. Examples include policies establishing criteria for the appropriate background for evaluators, providing for training on the appropriate roles for evaluators and managers, and safeguarding the independent role of evaluators through "whistleblower" protections.

6. Process and methods: An evaluation policy that belongs in this category establishes acceptable methods for evaluation, standards for methodological rigor, standards for technical quality, or standards for credible evidence.

7. Utilization: An evaluation policy that belongs in this category addresses making evaluation data or evaluation results accessible to external audiences, or the use of evaluation results generally.

8. Meta policy: An evaluation policy that belongs in this category provides for oversight or coordination of evaluation activities organization-wide. Examples include policies on periodic assessment of evaluation activities, policies on coordination of evaluation activities and policies whether vertically, across levels of the organizational hierarchy or horizontally, across peer sub-units of the organization, and policies related to the establishment of evaluation clearinghouses for sharing evaluation information.

Values dimension: aspect of evaluation policies related to what is important for the way in which evaluation is conducted or the way in which results are used, for example a concern for social justice, for evaluation quality, or for any of the principles contained in the AEA Guiding Principles or the JCEES standards for evaluation practice.

Using the ratings

Recall that the initial raw data containing 920 brainstormed evaluation policy ideas was sampled for range to obtain the broadest conceptual mix of statements for sorting. This was done in order to obtain as comprehensive an evaluation policy taxonomy as possible from these data. Statements were sorted and a cluster solution selected, and clusters summarized. At the same time, participant rating data was collected on two scales, “merit” and “feasibility”. To preserve comprehensiveness, rating inputs from study participants are not used to eliminate statements or categories here, so that even the lowest-rated statements and clusters are retained in the final taxonomy. However, ratings may be usefully applied by dividing the set of evaluation policy ideas in quadrants, from lowest rated quartile to the highest rated quartile. Users of the taxonomy interested in developing organizational evaluation policies may use the ratings (provided in Appendix E) to identify groups of evaluation policies or categories that might be addressed first, second and so on, according to their relative ratings. Ratings for all statements are provided in Appendix E. A listing of the twenty statements with the highest combined merit and feasibility ratings is provided in Appendix G.

Alternate taxonomy

In any classification scheme, reliable classification criteria and mutually exclusive categories are vital for inference (Krippendorf 1980, Weber 1990, Stemler 2001). One difficulty in reliable classification of policies is that they are inherently complex, and have many different aspects by which they might be grouped, such as benefits, costs, substantive aspects and financial aspects (Greenberg 1977). According to Mitchell (1985), this is why attempts to generate empirical taxonomies by examining a set of policies and grouping them by their observable characteristics tend to fail. Mitchell suggests the only way to achieve an exclusive set of policy categories is to generate a theoretical taxonomy, organized according to a set of analytical concepts or variables capable of describing and accounting for all possible policy variations, whether empirically observed or not (Mitchell 1985, p.11). Mitchell describes approaches to theoretical policy taxonomies which categorize policies based on 1) the fundamental social values they embody, such as democratic legitimacy, organizational efficiency and neutral technical competence

(Garms, Guthrie, Pierce 1978); 2) economic consequences, such as distributive, redistributive and regulatory (Lowi 1964) and 3) basic control mechanisms, such as resource allocation, rule making and the articulation of ideological beliefs (Mitchell and Iannaccone 1980).

In the present study, the results do not support a taxonomy with perfectly non-overlapping categories. In fact, the individual sorting schemes are fairly weakly correlated to the aggregate sorting scheme represented on the concept map compared to most concept mapping studies. Spanning analyses show that many of the clusters on the concept map contain statements perceived by participants as connected to statements in other, sometimes quite clusters. This is because different sorters grouped evaluation policy ideas according to different dimensions. A two-dimensional interpretation, in which all policies are assumed to exist on the same plane, forces an interpretation which ignores these additional dimensions. A deeper examination of the individual sets of participant sorting inputs behind the concept map shows why. The set of sorting inputs most highly correlated with the "average" groupings of statements on the aggregate concept map reveals categories which correspond to policy jurisdictions, with category labels such as "AEA/other professional organization [evaluation policies]", and "government evaluation policies". The second most correlated or "typical" individual set of sorting inputs contains categories which correspond to standards, guidelines or rules, with category labels such as "ethical standards [for evaluation practice]", "[rules about] evaluation procedures". The least typical individual set of sorting inputs contains categories with labels which correspond to goals or desired outcomes, such as "advancing evaluation", and "evaluation for improvement". The second least typical individual set of sorting inputs contains values-related categories such as "increase transparency, access and justice".

Synthesizing these insights from the data with insights from previous policy taxonomy frameworks from the literature, a reexamination of clusters and cluster names on the concept map suggests they represent a mixture of different **types** of categories, including particularly the dimensions of guiding values and policy mechanisms. Many clusters are more strongly joined by a value, such as “justice”, “usefulness” or “quality”, while others are more unified by a policy mechanism, such as the “principle”, the “universal standard” or the “guideline”. Yet it is possible to identify both values and policy mechanisms within every cluster. The map below shows the breakdown of a two-dimensional interpretation in which all policies must be assumed to exist on a single plane. Values dimensions are underlined in green, while mechanisms dimensions are *italicized* in red.

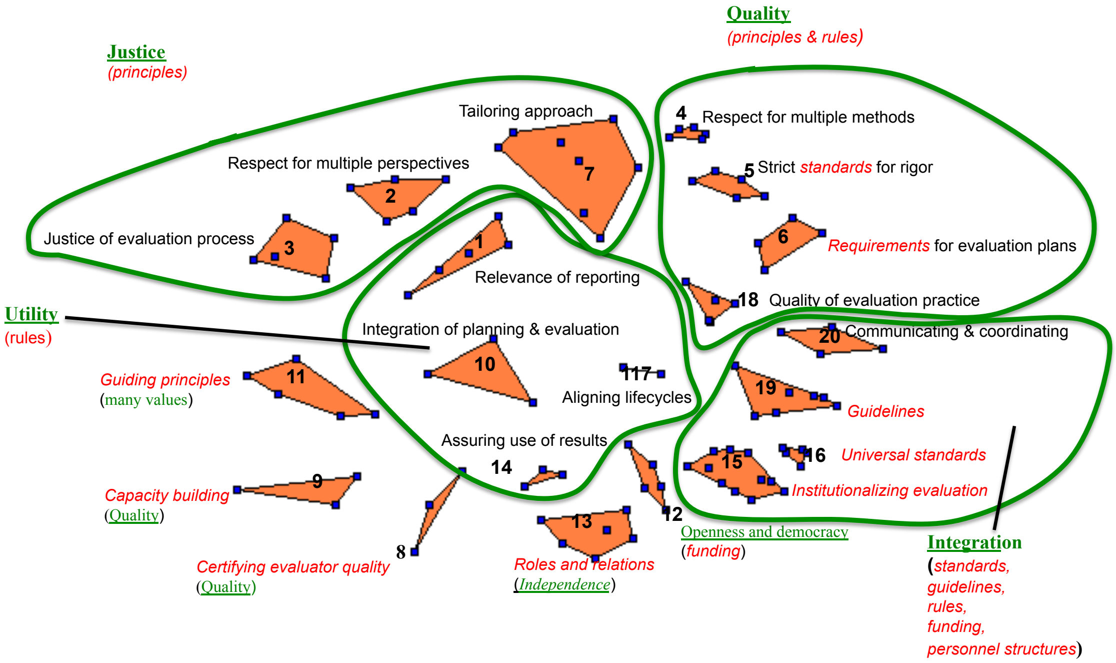

Figure 19: Values and mechanisms dimensions, cluster map

With the concept map as its foundation, a taxonomic framework incorporating the dimensions of value and mechanism appears below in Figure 20. Here, each column represents a specific value dimension and each row represents a policy mechanism, with associated concept map cluster numbers in parentheses. Values columns are presented from left to right on a continuum from most situation-specific to most standardized, beginning with populist values and moving through pragmatic concerns to theory-driven rigor to broader standardization of practice (see Datta 2011). Mechanism rows are presented from top to bottom on a continuum from most obvious to most subtle, beginning with the three control mechanisms corresponding to those of Mitchell and Iannaccone (1980) and then adding a fourth mechanism of role structures from the concept map (see also Datta 2009, Grob 2003).

Two exceptionally diverse clusters, 11 “guiding principles” and 15 “institutionalizing evaluation”, whose boundaries do not correspond neatly to distinctions between values or mechanisms are noted with an asterix.* In those cases, the identification number of specific statements associated with that particular mechanism row or value column are noted directly below the asterix. (Note that statements 36 and 61 are the only two statements allowed to straddle columns, since each represents a bundle of standards containing several different values.)

Table 7: Evaluation policy taxonomy: values by mechanisms

	justice (2,3,7,11*,12) *sts. 22, 61	utility (1,10, 11*,14,17) *sts.15,50, 36, 61	quality (4,5,6, 7,8,9, 11*,18) *sts.36,61	integration (15,16,19, 20)
statements of principle (11)	*statements of principle regarding justice of evaluation process or results*	*statements of principle regarding utility or use of evaluation results*	*statements of principle regarding quality of evaluation*	*statements of principle regarding integration of evaluation activities across units or among levels of a system*
standards, guidelines, rules (8, 15*, 16,19) *sts. 4, 9, 10, 23, 57, 88	*standards, guidelines or rules intended to promote justice of evaluation process or results*	*standards, guidelines or rules intended to promote utility or use of evaluation results*	*standards, guidelines or rules intended to promote quality of evaluation*	*standards, guidelines or rules intended to promote integration of evaluation activities across units or among levels of a system*
resourcing (9, 15*) *sts. 52, 53	*resourcing intended to promote justice of evaluation process or results*	*resourcing intended to promote utility or use of evaluation results*	*resourcing intended to promote quality of evaluation*	*resourcing intended to promote integration of evaluation activities across units or among levels of a system*
role structures (12, 13, 15*) *sts. 6, 26, 30	*role structures designed to support justice of evaluation process or results*	*role structures designed to support utility or use of evaluation results*	*role structures designed to support quality of evaluation*	*role structures designed to support integration of evaluation activities across units or among levels of a system*

Using the classification system proposed above, an evaluator could assist an organization in taking stock of its evaluation policies by examining relevant sources and populating the grid. The evaluator could then help the organization examine the grid for alignment of evaluation policies with organizational values, and for gaps and dissonances in the overall set of evaluation policies. The organization could then use the results to form a plan for developing its evaluation policies.

Needs assessment and inventory instructions

Following is a suggested step-by-step protocol for evaluators working in organizational contexts conducting an evaluation policy needs assessment and inventory. Note that instead of beginning with a set of standard areas evaluation policy should address, this process leaves it to the local working group to develop, through an initial needs assessment and through the inventory process itself, the areas where evaluation policy is in force and where it is needed.

Step one: Assemble a working group. The group should include key organization staff and might include a small number of external stakeholders, but should be small enough to make decisions without major difficulty.

Step two: Conduct an evaluation policy needs assessment to identify perceived problems related to evaluation in the organization. Generate a short list of goals the working group hopes to accomplish by changing or developing the organization's evaluation policies.

Step three: To find evaluation policies, search out, examine and document all of the organization's policies that might influence how evaluation is planned, conducted and used. To find formal evaluation policies generated within the organization itself, begin by looking for an organizational charter that might articulate overall policies or general principles. Look at staff job descriptions, for any written guidance to staff and/or grantees on how to evaluate. Look at reporting mechanisms and structures. Look for incentives related to evaluation (whether positive or negative). Look in the organization's budget. If the organization makes grants to other organizations, look at "request for proposal" language. To find implicit evaluation policies—so-called “rules of thumb” or unspoken norms about how evaluation is planned, conducted, or used, gather evaluation reports for the past few years—what repeated practices show up? Look also at published speeches of organization’s leaders which may contain expressions of goals related to evaluation. If possible, interview organization staff in key positions to find out how the “lived policy” may differ from formal, written evaluation policies. To get a clear sense of the context, ask about external evaluation policies or other forces that may be influencing the organization’s evaluation practices or use of evaluation results. Try to identify the interests of all stakeholders who might be affected by changes in evaluation policies. If important evaluation policies are set by entities outside the organization, include them.

Step Four: Examine the assembled set of evaluation policies for value alignment, policy gaps and conflicts. With the working group, do a visual inspection of each of the filled-in inventory template. Are clear evaluation policies lacking where they are needed? Are there evaluation policies in force that are causing problems? In examining the set of policies in terms of their informing values, do these values align with the organization’s core values? Do the values represented in the organization’s evaluation policies seem harmonious, or do they appear to be in conflict with one another? Is there a mix of policy mechanisms in use, or does the organization rely exclusively on one or two policy mechanisms, with evaluation policies all appearing in one or two rows of the grid?

Step five: Synthesize findings from step four with the working group’s initial set of problems and goals. Identify areas where the working group’s original goals and inventory findings intersect. Where the inventory has identified gaps or conflicts not included in the working group's initial set of goals, make recommendations for amending the list of problems and goals. Once the set of problems and goals is finalized, for each item, note any barriers to change, such as entrenched internal interests or external pressures. Generate ideas about how to overcome barriers. Using ratings inputs, make recommendations for addressing the most import and feasible evaluation policy issues first.

<u>Step six:</u> With input from the organization's middle managers and rank and file staff who will be charged with carrying out policies, help the organization make a multi-year plan for phasing in changes.

Reflections on how well concept mapping worked in this study

The use of concept mapping in this study made possible the inclusion of a large number of participants across large geographic distances in a short period of time, and the freedom of anonymity. As a result, the data probably contain a greater breadth of ideas than the results of single-theorist or leadership committee processes. Unlike the "Roadmap" process, which presented AEA members with an already completed statement of evaluation policy principles and asked them to react to it, the concept mapping survey in this study provided a simple definition of evaluation policy and an open-ended question to allow for broad brainstorming outside the box. Unlike political processes within AEA, concept mapping here provided a fairly open, systematic and transparent process for measuring the degree of consensus around the definition of evaluation policy, the salient types, and views on which specific evaluation policies should and could be implemented.

One inherent trade-off of concept mapping is that an aggregate sort result is gained while individual sorting frameworks are lost. The “slicing up” and averaging of participant sorting inputs destroys the coherence and internal consistency of each sorter’s individual classification scheme. At the moment sorting data is entered into the mapping process, participants’ precise classification criteria, or reasons for sorting statements as they did, remain unknown, and can only be inferred from their individual sorting inputs. The aggregation and translation of sorting inputs through MDS into a map in x,y space results in some groupings that are sensible and some that may not be. Because of this, the method calls for careful interpretation to assure meaningful results. Fortunately, concept mapping allows for a grounded interpretation process, in which the researcher asks members of the group to help make sense of the aggregate, cluster solution, restoring a grounded coherence. However, because the researcher did not seek participant interpretation here, researcher interpretations shape the results more than in some other concept mapping projects. Involvement of participants in these interpretive processes or even post-hoc member checking would have reduced the influence of the researcher and improved validity in this study.

Reflections on the utility of results

These results have expanded and refined the construct of evaluation policy so as to better articulate the domain, and have added to a prior theoretical taxonomy new dimensions, inclusion rules and example policies. Reliability testing is a needed next step. Reflections on specific components of the study follow.

Because so much of the data is discarded before the sorting stage of concept mapping in this study (820 of 920 evaluation policy statements), researcher decisions made in the synthesis and reduction process have the potential to influence the results by determining the sample of statements which will be used for the final map. One year after the initial coding and reduction of the statement set (initially done in summer 2009), the researcher reexamined coding decisions. The resulting reflections on the coding and reduction process appear in Chapter 4, Results, "comparing the original statement set to the final, reduced set." The Results chapter and associated appendices include an extensive audit trail on coding and reduction decisions and results, allowing external dependability audits of the coding and reduction process, and confirmability audits of the results (see Tashakkori and Teddlie 1998; also Lincoln and Guba 1985).

The main utility of the ratings results in this study is to help prioritize evaluation policies, possibly for organizational evaluation policy development, for advocacy, or for theory development and testing. Possibly due to a lack of variation in participant characteristics, rating results show little evidence of distinct sub-group voices on the issue of which policy ideas have the most merit. The only exception is that those who identify their primary work activity as "evaluation" appear to differ from those who identify their primary work activity as "research" in their views about which evaluation policies are more feasible. This suggests a fresh sample of raters with a higher proportion of researchers might yield different average cluster feasibility ratings values, but not necessarily different average cluster merit ratings values. This might render a meaningful difference between the pattern of average cluster feasibility and average cluster merit ratings, such as was not found here. However, in these results, there are a few exceptional clusters for which average merit and average feasibility ratings diverge somewhat. The divergences in relative average ratings for these few clusters suggest that raters were able, at least to some extent, to distinguish their view of the merit of ideas (“should be implemented”) from their view of the feasibility of ideas (“could be implemented”).

Reliability testing is needed to see if evaluation policies can be reliably classified using the inclusion rules proposed above. The categories corresponding to the cells of the “values-by-mechanisms” taxonomy and the step-by-step instructions for using the grid should also undergo extensive testing to see if evaluation policies can be reliably and usefully classified using this framework. Assuming a reliable version of classification criteria for both the could be developed, the scheme could then be used by organizations conducting inventories for purposes of evaluation policy development, by researchers developing and testing theories of evaluation policy, and in the eventual construction of an evaluation policy clearinghouse.

Appendix A-Original brainstorming statement set (N=920)

ID#	Statement
1	set aside funds to be spent on the networking of or meetings for evaluators working on similar projects to share ideas,methods and results near the beginning, middle and ends of programs (separate from the funds needed to do the evaluation work for the client program).
2	Require that 10-15% be spent on evaluation so that enough resources are available for meaningful evaluation efforts.
3	encourage the establishment of sustainable evaluation systems for grants that last 3 or more years so that recipients are encouraged to continue evaluation after federal funding ends in order to encourage ongoing data collection, analysis and reporting (perhaps with evaluation stipends that continue after the other grant funds end so that recipients have incentives to keep the data flowing).
4	the application of consistent methods for similar types of evaluations that allow for meta analysis of results for more in-depth learning.
5	federally funded programs must perform some type of program monitoring and evaluation
6	careful planning including consideration of available financial and human resources to determine when the evaluation of policies will take place
7	Design evaluation according to the project budget
8	establish measurable criteria by which the success of the policies will be evaluated in advance
9	Conclusions should allow for majority and minority opinions based on various stakeholders review and perspectives of the results.
10	Ideological philosophies should be noted as underlying assumptions in determining program theory or logic.
11	All aspects of a program's logic or theory should be considered appropriate for evaluation, as determined by the expected utilization of the results.
12	Recognition that "program failure" or "no discernable effect," as determined by an evaluation, is constructive feedback for program redirection and does not necessarily diminish the original rationale for its undertaking.
13	External evaluation is no more valid than "internal;" the rigor by which the design and implementation of the evaluation is done determines its validity.
14	Recognize that everyone is an evaluator, that "professional" evaluators only provide more sophisticated advice on approaches, methods and tools.
15	Allows for multiple methods and tools.
16	Allows for multiple approaches to evaluation, i.e., experimental, quasi-experimental or non-experimental (exploratory or in-depth case study)
17	Developmental, i.e., allows for the redirection of any evaluation as circumstances change and not simply a gap analysis from pre-determined targets.
18	Participatory, i.e., those participating in any evaluation are not simply asked questions to respond to, but have input as to the questions to be asked!

19	utlization of findings.
20	a clear timeline
21	understand the "end users" for the evaluation results.
22	the use of trained, independent evaluators who are provided with sufficient resources to execute comprehensive, multi-method evaluation designs.
23	use of multiple data sources or methods in conducting evaluations of student learning outcomes
24	the employment of rigorous research methods conducted by outside, external evaluators
25	how evaluation information is utilized during the program implementation, if at all
26	the federal government and all other funders must recognize that programs can not (and should not be expected to) achieve and report long-term, sustainable change after a few short months of startup and implementation. Unrealistic expectations for change tied to funding create climates where programs focus on telling the funder what they need/want to hear instead of working towards true program impact, which usually happens more slowly.
27	a basic level of data and evaluation should be expected of all programs, but a program's evaluability should be considered before expecting use of more expensive/extensive outcome or impact evaluation designs, especially those involving control and comparison groups.
28	A recognition and acceptance that many research designs are appropriate and valid for evaluation, and that while experimental designs are the "gold standard", experimental designs are not the only appropriate and valid means of discovery.
29	evaluation questions and program goals that are: 1) useful to the funder, the program, AND the community served, and 2) realistic given the duration, intensity, and context of the program
30	Increasing the relevance of evaluation to program managers, staff, and receipients by incorporating evalution activities throughout program implementation as a source of information for refining implementation in vivo (rather than evaluation occurring solely, or primarily, as a post-hoc activity)
31	evaluation should be a standard part of program development and implementation, included from the very beginning of program selection and design... from initial formative evaluation through outcome and impact evaluations
32	A means to detect and decrease symbolic or pro-forma evaluations that are conducted solely to adhere in a minimum way to a funders requirements for evaluation
33	evaluations should not be conducted by program managers working alone; program managers should have access to a cadre of trained evaluators to help guide the work
34	all federally funded program should include funds and requirements for evaluation
35	context and implementation is equally as important as outcomes
36	evaluation should use methods appropriate to the decisions that will be made about specific projects and programs - from formative evaluation during the development of initiatives, to qualitative studies exploring process, to mixed methods that captures outcomes.
37	staff must be culturally competent involving population served

38	that evidence of coercion through the exercise of power, influence, or resources be reported and the person(s) doing so be protected under Federal Whistle Blowers Laws.
39	that every effort is expended to capture contextual factors that impinge upon programmatic participation.
40	that costs to impact upon quality of life for all U.S. residents be articulated and considered paramount.
41	that the goals of the program must reflect the ethical grounding that the Nation attests to have with respect to fostering inclusion, maintaining human dignity, and empowering those who have been historically marginalized.
42	that the evaluator(s) must be able to substantiate cultural familiarity, sensitivity, and responsivity with those who are beneficiaries of the program.
43	that measures undertaken for collecting data germane to evaluating programs in all phases be appropriate in terms of reliability and validity for the respondents.
44	that evaluations be undertaken with the cultural lens (lenses) appropriate to derive an accurate basis upon which decision making may occur.
45	that evaluation rigor/design/methods should reflect the evaluation goals/uses, scope of the project/program, and available respurces/expertise.
46	funding for periodic evaluations should be required in federal programs authorized by Congress.
47	the recognition that some programs will require a mix of evaluation methodologies.
48	adherence to the evaluation standards developed by the Joint Committee on Evaluation Standards
49	no one policy fits all programs or situations
50	Methods for ensuring independence of evaluators
51	transparency throughout
52	openess, completeness, and honesty in reporting methods and results
53	A commitment to building evaluation capacity and an evaluation culture
54	ethical principles
55	Whatever rules and principles are adopted at the Federal level, they should be devised in such a way as to set appropriate standards for evaluations at other levels of government, such as states and localities.
56	I think there should be a policy that evaluations need to be conducted by individuals who have the training, terperament, and experience to do so.
57	There should be guidelines about aligning the methods and procedures of an evaluation with the information needs of the project and the rigorousness of the findings. While experimental designs remain the gold standard in program evaluation, there are other approaches that may be more suitable, given certain constraints and limitations.
58	be specific to the population served and culturally competent at a broad level.
59	how to incorporate with state level evaluation.
60	a policy that directly assesses innovation in terms of the level of risk associated with the research.
61	Encouragement of the use of multiple and appropriate methods.

62	The expectation that the evaluation respond to the needs of an array of stakeholders, including both federal program officials, project personnel, and those on whose behalf the project has been undertaken.
63	Reference to an established statement of evaluation principles.
64	that the full ADDIE process be conducted when creating or revising training
65	(1) flexible methods/methodologies depending on the questions that drive the evaluation are acceptable if conducted rigorously, (2) minimum (~15%) requirement in all budgets for evaluations (includes formative & summative work); (3) measurement of outcomes/impacts appropriately linked to time-lines of projects (i.e., no expectation for long-term impacts for projects that have only 1 to 3 years to design, implement, and evaluate); (4) well-designed program logic models (with evaluation directly linked to these; (5) summative evaluations should document how formative evaluation information has been utilized in the project
66	consideration for funding ramifications--for hiring an external evaluator to the staff time required to participate in or manage an internal evaluation
67	value for qualitative and quantitative methods as well as different levels of rigor depending on the goals of the evaluation
68	evaluation is included from the beginning--developing a logic model or theory of change to guide the evaluation design
69	methods and questions that are useful for the funder and the grantee
70	funding shoudl include realistic outcome evaluation
71	evaluation should be embedded in all program development
72	That evaluators go through a vetting process in order to be paid directly by the funder, not through the funded.
73	Articulation that evaluation provides information for decision makers and thus is part of program and policy management. Funding to pay for evaluation should come from management budgets, not program delivery budgets.
74	Training required for federal, state and program managers in how to manage external evaluation, including what evaluation is and what constitutes effective evaluation work.
75	one that values the levels of evaluation - at least local, state and federal - how they are linked, how they differ by function, and how they are all needed to contribute to a comprehensive evaluation of complex systems change.
76	Require all federally funded evaluations, many of which get buried, to be made public, or an extracted version
77	There should be an attempt by the Feds to publish evaluation results around specific topic areas to identify the state of the art and gaps in knowledge
78	Support and nurture up and coming evaluators, through fellowships or sabatticals at different agencies, co-authorship, etc.
79	Find ways to detect and decrease the amount of symbolic evaluation that is occurring, perhaps through reviews of current evaluation programs at all Federal, State and Local agencies and programs
80	Requirement that international programs, such as HIV prevention programs and others, are systematically evaluated

81	A committment to building evaluation capacity and an evaluation culture
82	the utilization of a theoretical basis (research and evidenced-based framework)that aligns with the evaluation site and its evaluation objectives and programmatic needs.
83	one that ensures the inclusion of diverse stakeholders and their perspectives as a foundational principle of evaluation, particularly enlisting their feedback and response prior to the dissemination of evaluation results.
84	one that addresses evaluative approaches that affirms and offers evaluation training on applying culturally responsive evaluation methodologies.
85	the evaluator should be trustworthy and competent.
86	Methods and procedures should align appropriately with the focus of the evaluation; multiple perspectives and approaches add to the strength of understanding a phenomenom
87	Evaluators should be affiliated in some way with professional standards or guiding principles (such as AEA membership, certification, etc.)
88	to require basic process evaluation and accountability data collection in all funded programs. Require rigorous outcome evaluation only when adequate funding and prior performance indicate "evaluability" is present.
89	Evaluation happen at all levels when program funds can support the evaluation. The evaluation must focus on program development with outcome data being collected at a future date, determined to be reasonable by the program, audience and anticipated outcomes (all of which should be identified at beginning of program development). Evaluation results should be made available to peers and stakeholders.
90	that local evaluators in demonstration programs be independet of the recipients of the grant, with separate funding stream to the evaluators.
91	that third party evaluation contractors and protocols be in place at the begininning of large multi-site demonstration programs and cooperative agreements, rather than hired on several years into the program.
92	to require basic process evaluation and accountability data collection in all funded programs. Require rigorous outcome evaluation only when adequate funding and prior performance indicate "evaluability" is present.
93	development of evaluation questions useful for program improvement
94	participation in evaluation by program staff and other stakeholder
95	When 10-15% of program funds are allocated for evaluation, the evaluation should include both formative and summative methodologies as well as meta-evaluation to assure responsiveness and responsibility to all program stakeholders.
96	evaluators should be certified
97	the over-reliance on experimental methods as the only acceptable methodology.
98	implementation evaluation strongly encouraged in addition to time (typically 2 or more years) for programs to be in operation before high-stakes outcome data are required.
99	Voice & space should be given to constituents/beneficiaries/end-users throughout the evaluation process, beginning with design.

100	Evaluation design & methods should suit the purpose and context of the evaluation; not the other way around!
101	Evaluation designs & methods should allow for different levels of complexities in change theories (simple, complicated, complex, chaotic; see Cynefin Framework)
102	Diversity impact on policy formulation and effects`
103	outcomes are context dependent and in many cases cannot be standardized
104	there should be a mechanism by which federally funded programs track evaluation systems - with the goal of developing a set of evaluation systems that - when applied appropriately to similar programs in different settings - comparisons can be made
105	comprehensiveness
106	the excessive reliance on indicators (counts) to the exclusion of a global project evaluation weakens the system of evaluation. This is exemplified in the PEPFAR HIV programs.
107	using the program evaluation standards developed by the Joint Committee
108	guiding principles which include ethical guidelines, and allows mixed methods
109	guiding principles which include ethical guidelines
110	encouraging the use of mixed methods and convergent findings as suggested by Heckman as alternatives or additions to RCT
111	to require an implementation evaluation along with time for programs to be operational before reporting high-stakes outcome data.
112	Value-Added Assessment
113	one that adds to the body of research ON evaluation.
114	informed by the most current knowledge about evaluation theory and practice
115	is in line with the Guiding Principles for Evaluators and/or program evaluations standards
116	allows for mixed methods
117	one that provides flexibility regarding the context in which a program or agency operates.
118	a higher concentration of evaluations that use a mixed-methods approach for more comprehensive findings
119	Evaluations should be meta-evaluated to ensure credible and valid evaluation findings
120	That the needs of the evaluation and the questions that are asked drive the methods, and that all methods are equally valued according to the purpose of the evaluation
121	the recognition that certain types of programs, such as those in Extension agriculture, are difficult to track to determine long-term outcomes. As such, provisions should be made to accommodate this kind of situation.
122	issues of equity are central provisions of federal programs and should be an element in any evaluation study of a program that seeks to assure both access and outcomeson an equitable basis.
123	federally-funded evaluations should automatically provide access to data sets maintained by public entities, such as schools, state welfare departments, clinics, etc., with appropriate IRB oversight and assurances of confidentiality, anonymity, and security.

124	task order qualifications should be opened on a frequent and regular basis to include newer organizations into the mix. There is often a same-old, same-old feel to the regular recipients of non-compete evaluations.
125	program officers should help negotiate the appropriate resources allocated for evaluation with both the project and the external evaluator. This will help the feds find answers to questions raised by policy makers and legislators, while at the same time protecting the resources spread between program and evaluation.
126	small business set-asides should be honored and true partnerships between large and small research firms encouraged.
127	All methods are viable option (quant, qual, mixed, etc.)
128	evaluators are encouraged to engage in a variety of methods to capture the uses and impacts of programs and projects. No single method is prescribed, and creative approaches are encouraged as a way to expand both the methodologies available and the range of outcomes that can be studies.
129	that a minium percentage of the project be allocated for the evaluation.
130	that the evaluator have a minimum level of education or training in evaluation.
131	IF the the client provides data, then the client is responsible for the accuracy of the data.
132	evaluations can be conducted both externally and internally, but both must include conditions of independence and non-retaliation. Separate contracts should be included with major proposals to ensure that funds available for evaluation are not reallocated to program efforts prior to completion.
133	that any evaluation should use methods appropriate to the decisions that will be made about specific projects and programs -- from formative evaluation during the development of initiatives, to qualitative studies exploring process, to mixed methods (including RCTs) that capture outcomes. Appropriate to decisions is the critical phrase.
134	A minimum of 10%, prefereably 15%, of any grant-funded project should be committed to evaluation. Further, there should be some minimum standards for the evaluation---or the evaluator (for instance, that evaluation is their primary role).
135	require a % of funds for evaluation, evaluation to be done by under the direction of qualified evaluators.
136	require external evaluator
137	madoatory fidelity assessment
138	minimum of 10% of project funding
139	no requirements for randomized, controlled evaluation designs (sometimes it's just no feasible)
140	mandatory evaluations by independent third parties
141	anything related to education or quality of life
142	Balance of comprehensiveness/quality with simplicity; sometimes less is more
143	% of time/resources that should be used for evaluation related activities
144	Recognition that strong evaluation requires resources
145	requiring a logic model or equivalent of a program so that it's clear that evaluation is informed by some sense of program logic

146	requiring periodic evalautions of programs, but ensuring these span both process and outcomes
147	establishing some miniumum threshold/setaside of program funds for evaluation
148	evaluability assessment conducted prior to launching into full-blown evaluation
149	logic models developed for each program
150	evaluation recommendations be addressed
151	any program with federal funding should be evaluated by a neutral body
152	Rules about conflict of interests between evaluators and agencies being evaluated.
153	For commonly measured outcomes, standardized measures of these outcomes should be established and required.
154	Standards should be developed for acceptable evaluation studies.
155	Federally-funded programs must include evaluation.
156	1. Develop evaluation competencies that would define professional preparation skills, standardize evaluation language, and develop competency-instruction workforce education and training. 2. Requirement for federally funded programs to include evaluation. 3. Evaluation data should feed back into program management. 4. Attention to mixed methods; respect for qualitative, quantitative, and participatory evaluation. 5. Attention to evaluation use - how results will be used for program improvement.
157	Developing and updating evaluation plans for each significant program.
158	at the marco federal level, there needs to be a movement to standardize the evaluation language. So that the evaluation work funded from one federal agency can be compared with the work from and other agency. Also, when and if there are professional-wide adopted professional competencies for evaluators with some degree of common preparation standards, there could be some possibility of being recognition as a "professional" class by the feds.
159	develop evaluation competencies that would define professional preparation skills, standardize evaluation language, and develop competency-instruction workforce education and training.
160	Requirement for federally funded programs to include evaluation
161	Valuing mixed methods
162	All proposed evaluation studies to have clear terms of reference including the purpose, objectives and scope of the evaluation, policy context, stakeholders, intended audience
163	the need to scan existing literature in the area at an early stage or before commencing an evaluation
164	ethical standards
165	the use of meta evaluation to asess the quality of evaluations
166	a database of evaluations would be created for access by the public.
167	one which allows for consideraation of substantive equivalence for anaylsis of course work done abroad that does not fit easily into the U.S. system (differing grading ssystems, credit allotment, and examination systems)
168	evaluators need to have formal education.
169	the evaluation process has to include mixed methods.

170	one which contains provisions for both General Education and Major area course work evaluation
171	funding are available to support the evaluation process
172	adequate funding for evaluation whenever new programs are funded or funding for existing programs is renewed.
173	that program funding contain adequate resources to produce robust evaluations.
174	a requirement that new programs be given time to properly form before summative evaluation (perhaps through requiring an evaluability assessment).
175	that federal (and other government entity) users of evaluation will share findings in a timely manner with the programs being evaluated.
176	clear expectation that evaluation is not completed strictly to inform policy, but to achieve other summative and formative aims.
177	as per the AEA standards, respect for the rights of the evaluand.
178	rejection of "one size fits all" evaluation, especially the over-reliance on experimental methods as the only acceptable method.
179	A standard for transparency and absence of political interference with evaluation findings.
180	evaluations must be conducted by experienced reseachers with appropriate educational credentials.
181	Timing of evaluation should be in sync with program design and planning.
182	The use of culturally sensitive practices by evaluators in carrying out evaluations on public program.
183	The Congressional Research Service should receive copies of all policy and program evaluations.
184	process evaluations and other formative evaluations need to be given priority with impact evaluations.
185	all program budgets should include adequate money for evaluation, even ifthe over-sight ofthe work itself is independent of the program.
186	all major policies and programs need to be evaluated, not just "targets," with appropriate funding to match the importance ofthe policy or the commitment of taxpayer resources.
187	appropriate resources be allocated to ensure evaluation at all stages of program
188	return on investment
189	evaluators follow a code of ethics
190	use of multi-methods
191	usefulness of findings.
192	independence of evaluators and program/policy advocates
193	accountability
194	that evaluators abide by a code of ethics
195	that methods should be appropriately adapted to suit the cultural or social setting
196	that findings should be true to the data gathered
197	that processes should be clearly articulated and transparent

198	that methods are rigorous regardless of paradigm
199	a policy that protects the rights and privacy of people who participate
200	an emphasis on formative as well as summative evauation, so the focus clearly includes oppotunities for improvrement of programs
201	do not require or design evaluation efforts that are more complicated than necessary ot that include exaggerated expectations
202	Related to ethical and human subjects interests, including pre-review of evaluation work not typically requiring IRB approval
203	one that describes adequate resources (including funding, time, expertise and knowledge)for planning and conducting evaluations.
204	the inclusion of the use of evaluation findings to make improvements in long-term projects/programs (e.g., the use of formative and continuous evaluation & utilization)
205	Evaluators are to work indepenently
206	external evaluator
207	consideration to addressing ethical and conflict of interests issues
208	I agree with culturally competent and utilization-focused evaluation that is disseminated for implementation of results
209	sufficient funding for evaluation, including evaluator time/fees/salary and participant incentives (if appropriate)
210	the use of multi-method research to triangulate the data and to learn not just what is happening but why, and how it is perceived by the people who are impacted
211	a measurable account of inclusivity/ cultural competence
212	A set of ethics Engages stakeholders and policy decision makers Addresses policy formulation, implementation as well as outcomes and impact Designed with use of evaluation findings in mind Uses evaluation standards No a priori requirement of a specific methodology (like current educational research policy) Guarantees of independence of evaluators with no repercussions for findings that do not match preconceived expectations and specific ideologies
213	evaluation plans that are clear, coherent, and purposeful.
214	usefulness of the evaluation findings.
215	that there should be no comprehensive set of evaluation policies
216	importantance of process evaluation to support the readiness of an outcome evaluation
217	a policy valuing partnership between academia and the surroudning community, with particular focus on low-resource communities.
218	a systematic way of ensuring evaluator independence.
219	reporting of evaluation findings in a manner that is jargon free, easy to understand by a lay audience and of high utility to the client
220	frequent sharing of cross-initiative evaluation findings
221	standardized programmatic data to be collected from each grantee
222	mandatory evaluation at the local (grantee) level
223	inclusion of relevant stakeholders, discussion, and deliberation
224	explicit identification of the standards for evaluation.

225	Attention to evaluation/research ethics.
226	Flexibility in evaluation design - opportunity to engage stakeholders in evaluation design and planning following the funding award.
227	culturally competent evaluation that helps states, local governments, and the non-profit sector to implement improvements.
228	information/results should be required to review and discuss with key stakeholders of program
229	budgeting for robust evaluation that is utilization-focused and culturally competent.
230	qua
231	Sufficient funding set aside in RFPs for evaluation activities (or a stated percentage of funds that will be used toward evaluation).
232	data
233	Include the evaluation design and requirements in the program design and requirements so that evaluation expectations and methods are built in from the beginning, not tacked on later.
234	Attention to mixed methods; respect for qualitative, quantitative, and participatory evaluation.
235	Attention to evaluation use - how results will be used for program improvement.
236	a shared database of methodology
237	programs, polices and practices should be required to conduct ongoing process evaluation and periodic outcome/impact evaluation
238	a protocal for relevancy or review
239	The flexibility to design evaluations based on the presenting issues and not be artificially restricted by silly government presecriptions on what constitutes sound research
240	resonable reporting of data, available to public
241	Accountability for every federal dollar spent
242	a standard of cultural competence to be upheld
243	an agreed upon set of ethics
244	culturally competent evaluation that is utilization oriented and disseminated to promote implementation of results.
244	culturally competent and utilization-focused evaluation that is disseminated for implementation of results.
246	evaluating programs for immigrants, homeless and migrants
247	The standard for the evaluation of social policy should be accoutable for advancing the goals of social justice.
248	I agree with a lot of statements below, such as "guidelines regarding the level of investment (cost) that might be required for various levels of evaluation rigor/methods. Currently, most evaluation embedded within program budgets is underfunded." or An accreditation process and licensing requirement for evaluators. or Requirement for Transparency in the evaluation process or Requirement of participation by end-users
249	Qualifications of evaluators Ethical treatment of individuals Adherence to some set of professional evaluation standards

250	qualified evaluators with appropriate credentials and experience should be contracted for external evaluation of programs and projects.
251	implementation and dissemination of results.
252	evaluation of policies designed to delay and impede government research especially those on sensitive topics (e.g., HIV, climate, violence, tobaccoetc)
253	examining the impact of research
254	adherence/compliance with Eval Standards (utility, feasibility, propriety, and accuracy) as well as competency of evaluator
255	value of non traditional forms of evaluation
256	That evaluation information is used to fine-tune promising practice, as well as test the efficacy of what is commonly described as evidence-based practices.
257	funding award includes dollars specifically for evaluation
258	consideration for required university IRB processes and the time that it sometimes takes to complete
259	should encourage and value systematic internal evaluation by providing funding, bonus points to applicants that have a demonstrable history of focus on evaluation internally.
260	options for low-cost evaluation strategies
261	rationale for the inclusion of specific evaluation protocol
262	opportunities for on-line training to ensure consistency
263	evaluation methods that are appropriate for the audience being served
264	Inspectors General (PCIE) Quality Standard for Inspections (see IGNET.GOV)
265	Guidelines to help evaluators avoid dual roles and responsibilities as program planners/implementors, and program evaluation.
266	evaluation data should feed back into program management
267	all hired evaluators demonstrate competency in evaluation and adhere to the AEA evaluation guidelines.
268	different dissemination methods will be used and that evaluation findings are documented in such a manner that ALL stakeholders can understand what the findings are.
269	providing more education to federal program managers about the multiple uses of evaluation.
270	An accreditation process and licensing requirement for evaluators.
271	the design of the evaluation should be appropriate to the research questions, employing methodologies that are rigorous and are a good fit (ethically and methodologically) with the program content and context.
272	trying to be cost-efficient by combining evaluations of multiple programs.
273	evaluations should include evaluation questions that are directly linked to project goals
274	an evaluation plan is established, documented and finalized within 60 days of funding and that the evaluation plan is updated annually.
275	Dissemination of data/info.
276	A requirement that when an evaluation team approach is adopted, the evaluation team include both content experts and profressional evaluators.

277	valuing a wide range of methods (qualitative & quantitative) and research questions (implementation & impact).
278	no study can be solely quantiative in nature.
279	viewing evaluation as part of program management that begins when the program begins and continues after the program ends to track longer-term impacts and outcomes.
280	Efficient budgeting
281	Establishment of a national repository of evaluation reports, organized by type of evaluand (e.g. program, policy, product etc.) sector, type of organiation, topic areas (e.g. child welfare services, teaching training programs, etc), and accessible to evaluators and program planners. Meta-evaluations could also be part of the collection. This would provide opportunities to accumulate evaluation "cases", moving the field beyond conducting single study evaluations which provide limited learning opportunities.
282	.... central data storage of eval data results that arew publicly and easily accessible
283	... ensure each RFP etc. includes an evaluation component with quantitative guidlines.
284	When funding is granted the organization/association/agency should be mandated and supported to implement or address the subsequent findings and/or recommendations from the evaluation.
285	and fit the program and
286	No comment
287	???
288	Rejection of evaluations that only include a QUANTITATIVE method of inquiry. Understanding that the linearity of logic models do not apply to all communities, all conceptualizations of problems, and all interventions.
289	Findings from the evaluation report should be made readily available to the general public, funding agency, and organization.
290	Grants cannot require the collection of any information that would be a violation of local, state or federal laws (such as FERPA requirements).
291	Adherence to high standards of professionalism
292	Qualified persons to design and conduct an evaluation that adheres to AEA guiding principles, is appropriate for the project/location/and questions to be answered, and follows a model generally accepted in the program evaluation community. Adequate funding should also be earmarked specifically for evaluation activities.
293	A policy on the qualifications of evaluators
294	A policy on how evaluations should be funded
295	A policy on ethical behaviors of evaluators
296	release of data to the public domain (i.e., to allow others to replicate or to use data for other research purposes).
297	funding that includes adequate follow-up to assess sustained or long-term program or policy effects.
298	a preference for multi-site, multi-method designs.
299	guidelines for the protection and ethical treatment of subjects.

300	Mixed methods evaluations: rather than relying strictly on RCTs or other forms of evidence considered to be rigorous, qualitative inquiry should be encouraged to illuminate the findings (success/failures) of evaluation studies
301	appropriate funding relative to the evaluation questions (and methods)
302	that evaluations should be evidence based
303	Evaluations should be utilizaiton-foucsed, with clear goals and strategies for dissemination, should speak to or inform the body of practice, and should speak to or inform the policy-level
304	clarifying for state, district, and school staff when FERPA does and does not apply to use of data.
305	the inclusion of both process and outcome evaluations.
306	compiling in a central location all evaluation reports conducted on federally funded evaluations that is open to evaluators and the general public.
307	Percent of funding for evaluation should be commensurate with the methodologies required/proposed.
308	If you have not already done so, see "Federal Evaluation Policy," a book written in 1974 by Wholey and others at the The Urban Institute's Program Evaluation Studies Group.
309	the independence of the evaluator to conduct the evaluation without funding agency interference.
310	that data will be made available to the public
311	consequences for professional misconduct
312	evaluators must be identified to evaluate whatever is being funded prior to disbursement of federal funds or shortly thereafter.
313	the evaluation design be appropriate for answering the evaluation question(s).
314	complex systems-based evaluation methods be used in complex adaptive systems.
315	Only the written English language should be used in studies.
316	that those responsible for the evaluation assert their biases upfront.
317	appropriate funding for the requested evaluation design
318	an awareness of the distinction between evaluation within the United States and that possible within developing countries.
319	requiring that a set minimum amount be required to be set aside for evaluation of any federally funded initiatives.
320	teaching students life & parenting skills and evaluating how they do on into they college years and beyond into family life.
321	A statement of personal bais for both evaluators and those who will interpret evaluations. I do not agree that evaluations are independent and objective, because they are performed by people and people are not independent and objective. Recognizing one's bias is a huge step in making the process transparent and informative for all at stake.
322	Mixed methods beyond RCTs be recognized as appropriate options - RCTs are not the appropriate "gold standard" in all situations.
323	Adherence to AEA guiding principles

324	can't think of anyting at the moment
325	can't think of anything at the moment
326	mixed methodology should be employed whenever possible.
327	all Federally funded projects should include process and outcome evaluation components to measure implementation and success.
328	translation of data collection instruments into languages other than English to ensure that all voices are heard
329	The decision-makers and types of decisions to be supported by the evaluation should be clear. The methods should be appropriate to the purpose, resources and intended completion date of the evaluation.
330	should have a minimum budget allocation for programs. Also restructure the way they do RFPs - they typically solicit evaluators to conduct methodology exercises (surveys, focus groups, etc.) rather than focus on a specific (and useful) research questions and indicators.
331	improving student achievement
332	the independent, objective and comprehensive nature of evaluation.
333	..all project proposals must include a detailed evaluation plan that includes process and outcome measures assessed throughout the program.
334	Rejection of evaluations that only include a qualitative method of inquiry.
335	when programs funded by the same rfp are being evaluated, get the evaluators in touch with one another
336	Evalution should be part of program design and implementation.
337	the objectives of evaluating programs and how the results will be used.
338	commitment to wait for the evaluation results before making a legislative decision about continuation of the program.
339	understanding that research agencies require special evaluation standards.
340	10% of program funding for evaluation.
341	a dedication to scientific rigor.
342	appropriate funding of evaluation for all programs.
343	Designate 10% of program funding for evaluation
344	Ethical guidelines regarding evaluation standards, application of evalaution approaches, use of data and evaluator actions.
345	the evaluation policy should provide adequate opportunity for stakeholder input that is lingusitically and culturally tailored to the population engaged in the evaluation activity
346	that evaluation by an outside evaluator is required as a condition of funding.
347	Type and level of education of the evaluator
348	adherence to the evaluation standards as developed by the Joint Committee at the Evaluation Center, Western Michigan University
349	adherence to AEA principles and ethics

350	federal grant review panels must include at least one evaluation professional to assure that the evaluation design is precise and appropriate, and that data collection and analysis methods will provide federal project officers with accurate information to determine the outcome/impact/performance of grant-funded projects (individually) and programs (collectively.)
351	...all agencies or organizations receiving government funding are required to have an appropriate evaluation plan included in the funding proposal.
352	evidence-informed decision making
353	Dissemination of evaluation will be tailored and fitted to the range of evaluation stakeholders. Dissemination methods (e.g. published reports, newsletters, electronic, media) are timely (e.g. within a set time time period).
354	Evaluation are part of program design and implementation
355	clear, concise statement of how the evaluation results will be used including stating what decisions will be made and what is the information / data point for making the decision (e.g. given finding x action y will be taken)
356	designated funding level for evalaution
357	review of prior research
358	10 percent of funding is designated for evaluation.
359	that a cost analysis (benefit, effectiveness, utility, and/or feasibility) be conducted.
360	that every program include an evaluation plan.
361	that every program provide an action plan for addressing evaluation findings.
362	that evaluation be a requisite component of all programs and projects funded
363	that evaluators should serve as key members of the planning body
364	that evaluation functions must be adequately funded
365	a reference to how each proposed evaluation relates, or not, to specific AEA standards. If the evaluator choses not to use these standards in the evaluation strategy, then a statement explaining/justifying this decision should be included. I do not agree that that such policies should include "points for background and experience of the evaluator" as stated by one of this survey's respondents. Perhaps "...one policy that should be included is that there be a statement, as well as a resume, of the evaluator's interest in the particular evaluation proposal along with a clear statement of how the evaluation results will be used and with what audiences.
366	Percent of budget allocated to evaluation should be realistic (10-14%)
367	Points for background and experience of the evaluator
368	evaluators be part of the planning process
369	a requirement that all new programs be evaluated from the time of establishment; a requirement that all existing programs be periodically evaluated for outcomes and impact within a specified period of time not to exceed 10 years.
370	a specific plan for disseminating the results to key program stakeholders and how the results will be used. Furthermore, those stakeholders should be identified at the beginning of the evaluation.
371	a statement of how the evaluation fulfilled the AEA's standards

372	understanding the evaluand's context which should subsequently guide the evaluation design, implementation, and reporting
373	Recognize that there is as much to learn by "failure" as "success" thus rewarding a complete evaluation that does not just report on the success and neglects discussion of the failure.
374	definitions of each indicator
375	adherence to AEA Guiding Principles and Joint Committee Evaluation Standards.
376	that evaluators should not be unduly pressured or subject to interference by agency/program managers in making decisions about sound, appropriate evaluation design, data collection, analysis, interpretation of results, reporting, and presentation of findings.
377	that the dataset be well documented and made publicly available.
378	a policy higher education budget cut.
379	a policy on ethics
380	Importance of process and outcome evaluation
381	evaluations should include program stakeholders to assure appropriate interpretation of results
382	a requirement that at least 10% of a project's budget should be for evaluation
383	evaluation that is appropriate for specific programs is encouraged, not "one-size-fits-all" evaluation
384	programs that get funding for evaluation should have to spend that money on evaluation
385	a program's stability/evaluability should be considered before expecting use of more expensive/extensive evaluation designs, especially those involving control and comparison groups.
386	the evaluation design should chosen with a full understanding of the context of the organization/program, not some a priori decision about what constitutes the "gold standard."
387	that any study using taxpayer money should have full transparency.
388	that evaluation results and science should guide program decision making; NOT politics, religion, or ideology.
389	evaluation of programs that are funded through federal grants
390	evaluation of educational programs
391	how the results will be shared with stakeholders and the public at large.
392	to focus on utilizing the results.
393	to ensure that human subjects are protected.
394	information to contextualize findings should be collected, and results should be disaggregated to inform policy in different settings.
395	OMB Circular guidance A-133, A-21 or A-122, A-110. Any federal regulations and guidance regarding the Government Results and Performance Act.
396	Stakeholders consultation
397	communication of evaluation results to public
398	periodic synthesis of evaluation findings for groups of programs
399	independent professional evaluation of funded programs.

400	to ensure that the results of a well-designed evaluation will be utilized for decision-making purposes.
401	Independent evaluation
402	the anticipated use of the policies, and an analysis of their cost-effectiveness
403	cultural competence
404	adherence to participant rights
405	to ensure that the evaluation is funded and external to the agency conducting the program.
406	that federal programmes include finacial provision for evaluation.
407	to make provision for the wide dissemination of the results and for free public access to the evaluations
408	evaluation results will be reviewed and used to inform decisions about continuation or expansion of programs
409	transparency of data and results
410	including evaluation planning with program design, so that an evaluation plan is agreed upon before the program begins
411	that evaluation is treated as an integral part of all programs
412	the use of evaluation results to inform federal decision-making
413	adhering to AEA evaluation standards for federal projects
414	the certification of federal evaluators for federal projects to be completed through AEA
415	the requirement for evaluation of federal projects
416	the amount of money that should be budgeted for evaluation of federal projects
417	there is are different types of evaluation purposes and approaches that must be aligned with available capacity including the resources of time and human will.
418	that evaluation rigor must align with the rigor of the program design.
419	the acknowledgment that culture extends beyond ethnicity, county of origin, language,etc. but that of academic disciplines (i.e., social work vs public health) and sectors (i.e., philanthropy vs social sectors)
420	ethical evaluation practice
421	stakeholder involvement (participatory evaluation) with a commitment to use results for ongoing improvement and/or action deemed important/beneficial by stakeholders
422	scoring rubrics including evaluation plans should be published in advance as a part of RFP application review.
423	evaluators should be included in writing RFPs for grants to focus on identified outcomes.
424	the effectiveness of each federal program should be assessed by an external evaluator.
425	ensure reasonable funding for evaluation and score proposals based on sufficient funding alloted to evaluation
426	ensure use of methods are guided by nature of research questions
427	clear communication of the role of evaluation in guiding continuous improvement
428	use of indicators to measure progress/impact that emerge from the experience of particpants

429	a rationale for the indicators used to measure impact
430	ensure that methods are appropriate to the purpose of the evaluation
431	evaluation initiates in advance of or in concert with implementation of programs or product development.
432	include stakeholders in the evaluation process
433	adheres to the guiding principles for evaluators as set forth by the AEA
434	protection of evaluation participants/subjects
435	compliance with human participants regulations
436	make an effort to listen to the weakest voices
437	to allow for flexibility of methods without valuing one method over another.
438	to make evaluation processes transparent and understandable to a lay audience.
439	to ensure that methods are appropriate for the evaluation's purpose.
440	One that enforces sound evaluation practices.
441	One that understands the social impact of the evaluand.
442	One that guards against misuse.
443	One that ensures use.
444	One that clearly communicates the intention of the evaluation process.
445	Give voice to those that otherwise might not be herd.
446	Value the different cultures being affected by the outcomes.
447	Respect the values of others.
448	Measurement of quality characteristics of how well services are delivered
449	Encourage mixed methods to determine if indicators have been met.
450	consistency with industry standards and best practices, including ethics and sound methodology
451	a commitment to openly sharing the results of the evaluation with the public
452	a commitment to utilization of evaluation results in future decision-making and action
453	a means to ensure input from all appropriate and impacted levels of the group or organization, not just the leadership
454	one that seeks to protect the contributions of those who may provide negative (politically unwelcome) facts or data, like whistleblower protection
455	ensuring availability of resources to enable a high quality, ethical evaluation.
456	inclusion of all stakeholders in the evaluation process.
457	use of multi-method approaches in data collection
458	one that allows evaluators to be flexible enough to truly address the goals of the project
459	transparency in identifying the information and its sources that guided the decision.
460	collecting accurate information on costs and consequences of alternatives.
461	a consistent methodology (based on a logic model) that defines, justifies and adequately measures key outcomes; defines, aligns and adequately measures critical formative assessments and information feedback loops; serves the reporting requirements of the external stakeholders and provides information that is perceived as valuable by the client.

462	the goal of the evaluation. Be it to evaluate the merit of a program, recommend changes to a program, or terminate a program.
463	ethical standards to be met by Evaluators in carrying out their work.
464	consideration of practice base evidence
465	use of results
466	Ethical practice
467	evidence-based decisionmaking.
468	egalitarianism.
469	recognition for the importance of autonomy at the local level develop an appropriate example for the particular context
470	purpose of the evaluation ought to be clear, and methodology/reporting/evaluation practices aligned with these purposes
471	alignment of federal evaluations with Joint Committee standards and AEA guiding principles
472	logic models or other clear statements of activities and expected outcomes
473	following ethical guidelines like those developed in AEA
474	use of appropriate measures that include qualitative as well as quantitative methods
475	inclusion of all stakeholders in the evaluation process
476	encouragement of implementation evaluation as well as outcome evaluation
477	Use of evaluation professionals
478	a requirement to review any recommendation enacted as a result of the evaluation at an appropriate interval to determine if they were successful.
479	taking a persective to cumalatively build theory over time about what works
480	evaluation should be done by evaluation professionals
481	transparency with respect to data, methods -- in other words, reproducibility
482	disclosure of supporting data
483	standards of methodology
484	standards of evidence
485	explicit criteria should be developed for identifying "promising" innovative practices
486	PART
487	an emphasis on professional ethics.
488	the requirement to work at the community level and any agency being funded need to meet minimum standards in order to carry out evaluation. There needs to be clear goals and an infrastructure to put in place so that agencies are accountable for what they say they will be doing.
489	an evalaution framework that includes stakeholder input with clear golas and accountability measures
490	emphasis on independent evaluation while also providing data collection and tracking capacity building for program sites
491	efforts to connect and collaborate across similar programs - consolidation of information to better inform any given field
492	qualitative elements as well as quantitative measures

493	fiscal accountability, including cost/benefit analysis whenever possible
494	Impact evaluation should be required for any RFP advertised.
495	Priority for comprehensive evaluations of programs based on realistic criteria
496	accountability for teachers that is not primarily based on student achievement
497	Emphasis on theory
498	Measurement of all aspects of function, not just expected outcomes. Otherwise, there will be too many unanticipated consequences
499	Assessments of shared understanding of program goals and mechanisms and that understanding changes over time
500	Identification of "feed-forward" mechanisms effectiveness
501	A focus on tracking all steps of the program development and impelmentation to meet goals
502	the ethical standards that govern the profession
503	That results are accepted, discussed and used by a committee of program and policy staff as well as academic experts.
504	The contract for the federal program evaluation should not be overseen solely by the program under evaluation. It should either be overseen by a branch or division that is over the entity being evaluated, or jointly with another division. This will eliminate the problem of contractors worrying that they won't be rated well by the COTR if they report adverse findings.
505	a recognition that scientific approaches are not practical or ethical for many funded programs.
506	a requirement for both external AND internal evaluation as well as the dissemination of evaluation findings to the general public and other programs.
507	evaluation funding should be a separate line from the program funding
508	requirement that evaluators on federal or large scale projects have had previous evaluation experiences and some coursework in evaluation theory
509	guidelines for internal evaluation and external evaluation
510	i'm not sure if just a set of rules or principles can capture a variety of decisions and actions all at the same time. perhaps what's needed for federal evaluations or any evaluations for that matter is that at the beginning of evaluation, there should be discussion and some consensus among stakeholders on what is the mission and purpose of the specific evaluation at hand and what rules or principles could serve achieving the agreed upon mission of the evaluation in the context.
511	widespread dissemination of findings
512	evaluators' experience and qualification
513	sustainable evaluation practices that are integrated into every program
514	transparency and full disclosure of all data and all findings
515	an emphasis on utilization and empowerment evaluation approaches.
516	claer guidelines for utilization of evaluation recommendations
517	that evaluation lessons learned and recommendations must be utilized and show evidence of being used to inform program development or the evolution of the program.
518	evaluators have demonstrated experience and education related to evaluation.

519	that multiple approaches be acceptable, depending on the nature of the evaluation questions.
520	that evaluations include both process and outcome components.
521	an external evaluator be used.
522	Evaluation lessons learned and recommendations must be incorporated into future programming and strategies.
523	Standards
524	adapting the standards to resources available in the communities or to methods acceptable in the populations.
525	that the population or program benefits from the evaluation
526	a principle statement that there is no one right way or one size fits all(internal or external, 5% to 30% of funding to evaluation, random or comparative design, formative, process,or outcome evaluation, qualitative or quantitative methods, etc.)
527	Basic training for evaluation literacy should be provided to all managers and staff on an annual basis.
528	clerly the following lists includes all my concerns
529	to document that the evalautor has specific training in evaluation methods rather than just research/statistics skills.
530	require that organizations explain and document how evaluation findings have been used by ALL stakeholders.
531	requires programs to use 10-15% of operating budget towards external evaluation
532	requirement for external evaluation
533	a requirement for evaluation on federaly funded projects by those trained in evaluation (rather than research) methods.
534	evaluations of the impacts of interventions must utilize research designs that support strong causal inference.
535	significant resources are devoted to evaluation, with the recognition that evaluation takes time
536	programs have targets and are held accountable; data is transparent; evaluations are conducted and results are used for quality improvement
537	a standard framework for evaluation such as CDC's framework Judgment of the merit or worth of a program, policy, strategy, event, or initiative Whenever appropriate, evaluators should include key stakeholders as contributors to the evaluation plan and implementation
538	one that helps determine what standards are appropriate.
539	the need for integration of the Program Evluation Standards. A firewall should be established between the evaluators and those who have a vested interets in the outcome of the evaluation.
540	the need for integration of the Program Evluation Standards.
541	a standard framework for evaluation such as CDC's framework
542	If multiple implementations of a program are to occur, a set of common outcomes and common measurements of those outcomes should be established.
543	Evaluations should be held to obtaining an 80% response rate - and funded so they can

544	A firewall should be established between the evaluators and those who have a vested interets in the outcome of the evaluation.
545	A logic model should be established with input from all stakeholders to guide the evaluation. It should be used to determine the emphases of the evaluation - how resources will be allocated - and it should be used to insure that the evalaution is not directed into other areas or outcomes while underway.
546	Any evaluation document should include guidance on how the evaluations results should be interpreted and used
547	Whenever appropriate, evaluators should include key stakeholders as contributors to the evaluation plan and implementation
548	There is no one gold standard methodology - evaluators should choose the methodology that best fits the evaluation question and intended use of the evaluation
549	Judgment of the merit or worth of a program, policy, strategy, event, or initiative
550	Guidelines for the conduct of evaluations
551	A clear delineation of when and under what circumstances evaluations should be undertaken
552	Professional standards for evaluations
553	A description of the nature and timing of formative and summative evaluations
554	Definition of program evaluation
555	collecting information from teachers and students about their experiences in schools
556	expectations for various standards and levels of evaluations
557	national framework that evaluators use to guide them
558	national guidance that all evaluators abide by
559	an evaluator or evaluation team (or someone skilled in methodology) should assist in development so that the data collection process is sound and appropriate.
560	outside evaluators should be used to ensure impartiality (in as much as impartiality is possible by an evaluator).
561	evaluation designs should include a mixture of quantitative and qualitative data collection tools (mixed-methods).
562	policies and decision-making on the continuation or re-authorization of programs should be based upon data collected from sound methodological designs.
563	that to the extent possible agencies and programs should share their measures and results to encourage the use of identical measures across programs/agencies where appropriate.
564	that constituents (grantees or others whose projects/programs are being evaluated) should have a voice early on in the evaluation design. Specifically they should be able to weigh in on, but not necessarily direct, the data collection methods and schedule, the range of data that will be gathered, the goals of the study and the products that will be developed through the study.
565	standards for ensuring that the evaluation method is appropripate to the project/program under study
566	Evaluation may be sponsored by the administrative system, but not 'bought' by it - evaluation is a public good.

567	Political, programme and institutional managers should commission evaluation in order to learn from practitioners about the complexities of social action.
568	Evaluation should be built into the program design with clearly stated evaluation work plan and allocated budget
569	Transparency and sharing of evaluation results with all stakeholders
570	sharing of findings is mandatory and a searchable site established to share results (to avoid publication and other biases against publishing program evaluation findings).
571	a variety of methods are welcomed, depending on the evaluation question(s) evaluation should be built into the program design evaluations should be carried out by external parties unless said external parties engages in increasing capacities for internal program evaluation
572	funding for evaluation commensurate with required evaluation activities.
573	when evaluations should be carried out.
574	shared database of evaluation reports for public programs
575	explicitly addressing the validity of evaluation findings - what are the threats to validity and how have those been addressed?
576	Use of consistent methods for assessing cost-effectiveness of programs
577	Evaluator ethics
578	inclusion of outcome and results indicators in all SOWs
579	the presence of an evaluator in the project design phase
580	the consistent style of construction of reports
581	Using local consultants as support evaluators to build capacity
582	formalised evaluation should be included from the earliest proposal and design phases and planned for appropriately from the perspectives of both resources and time
583	full disclosure of findings.
584	subject projection.
585	based on evidence-based literature that is shown to be effecitve for the populations of focus
586	Procedures for evaluation; transparency and fairness of evaluation
587	mixed method
588	a variety of methods are welcomed, depending on the evaluation question(s)
590	evaluation should be included from the design phase and budgeted for appropriately
591	that every federal program should consult with a team of experts in evaluation and create a plan for how evaluations of grant initiatives will be used to inform decision making
592	if expenditure of public funds support a program, evaluation for accountability of the expenditure on processes and program outputs, as well as the societal outcomes of the program should be incorpoated in program planning
593	human subjects protection oversight should be provided by a committee such as a QA/QI/Program Evaluation committee, separate from the IRB, as much of QA/QI/Program Evaluation does not qualify as research under the IRB definition.

594	for programs, evaluation should be planned at 7+or-3% of the program budget; lower if the evaluation is of lower complexity or importance with few or no residual products coming out of it, higher if the opposite
595	reporting of the technical and methodological elements of the evaluation should be separate from the findings and recommendations and should meet standards for professional documentation and written for a professional audience, with interpretations to support a lay audience's understanding
596	reporting of the evaluation findings and recommendations must be tailored to the specific audience that is expected to consume them
597	the evaluator should be chosen on the basis of skill and demonstrated competence rather than on political connections
598	the qualifications of the evaluator should match the research need of the study being initiated
599	the level of rigor of the evaluation should be positively related to the magnitude and criticality of the decisions to be made on the basis of results
600	the philosiphical biases of the evaluator should be clearly identified
601	interpretation of the data must acknowledge the data analysis used and be appropriate in the context of the evaluation
602	data analytic techniques employed must be fully described and replicable
603	data gathering must be appropriate to and directly relevant to the purpose of the evaluation
604	individual data gathered for one purpose cannot be used without affirmative permission for any other purpose
605	individual privacy is the default position, if any data are to be reported that could lead to individual identification that fact must be made available to the participant before data are gathered
606	participants in the evaluation must know why their data is important and how it will be handled
607	purpose of the evaluation must be clearly stated and included in statements about it
608	GPRA indicators should be reviewed at least every other year and funded programs leadership should be asked for input regarding relevant indicators of program success.
609	evaluation of clearly stated outcomes included in the program design. Outcomes do not have to be only participant change. Mixed methods for determining if goals of the program were reached and clear guidance on use of the evaluation....how it will be used.
610	relevant outcomes measures
611	Build evaluation into the overall program design. Evaluation design include consideration of how the evalution will be used and who will use the evaluation results. Evaluation findings should be presented in a manner that makes then useful and accessible to stakeholders at all levels including those involved in implementing programs. cultural diversity produces 'different ways of knowing' - these different ways of knowing should be incorporated at the highest levels of evaluation policy
612	applying deliberative democracy with a focus on true reciprocity to all evaluation and assessment procedures.

613	review of process, implementation and theories underlying complex organizations policy development and decisions at micro annd macro levels
614	the need to correlate program objectives with overall agency goals and define related, meaningful performance measures.
615	Build evaluation into the overall program design. Evaluation design include consideration of how the evalution will be used and who will use the evaluation results. Evaluation findings should be presented in a manner that makes then useful and accessible to stakeholders at all levels including those involved in implementing programs. cultural diversity produces 'different ways of knowing' - these different ways of knowing should be incorporated at the highest levels of evaluation policy
616	understand how to evaluate each program on a continuum from start through closeout over time.
617	include all stakeholders and mechanisms for feedback from them in the evaluation process.
618	tell us how you plan to demonstrate progress toward program goals.
619	build evaluation into the overall program design.
620	Participatory studies using emergent methods as a means of accessing answers to complex research questions and revealing subjugated knowledge. These techniques are useful for discovering knowledge that lies hidden, that is, difficult to tap into because it has not been part of the dominant culture or discourse.
621	Real world considerations when developing the evaluation methodology. Often-times, the emphasis is too much on the end result, without taking into account the characteristics of the target population (e.g. risk factors, transient nature, mandated reporting, etc.).
622	Evaluation should focus on examining how to quantify behavior or incorporate statistical methods to examine the relationship of non-quantitative variables.
623	Evaluation design include consideration of how the evalution will be used and who will use the evaluation results.
624	When designing the evaluation, a "benefit/cost" analysis of the proposed evaluation be conducted (at least at the conceptual level) to ensure that the evaluators and funders are demonstrating responsible stewardship of evaluation and participant resources.
625	Efforts be taken to make the evaluation process transparent to the public: easy access to information about the initiative being evaluated and about the evaluation
626	Evaluation findings should be presented in a manner that makes then useful and accessible to stakeholders at all levels including those involved in implementing programs.
627	Logic models be constructed to describe both the initiative and the evaluation of the initiative
628	Guidlines on dissemination processes of evaluation findings
629	Evaluation be done in ways that help identify means for making improvements in projects/initiatives so that outcomes are achieved to a greater extent
630	Evaluation as an expectation for accountability purposes
631	Recommendations on systems approaches to evaluation

632	Evaluation address both implementation and outcomes
633	Promotion of mixed methods approaches to evaluation
634	Standards of practice for evaluation conduct including qualifications of evaluators
635	Efforts are taken to productively involve people affected by the initiative/project in discussions about the evaluation: what questions will be addressed, what will be accepted as evidence, etc.
636	Guidelines for human subject protection/IRB requirements that make sense for low-risk evaluation studies
637	Evaluation as a tool for organizational innovation and development
638	Recommendations on integrating funding for evaluation into initiative budegt
639	Limitations of the evaluation methodology and the findings are included in the report
640	the evaluation thoroughly describes the situation in which the initiative is taking place
641	The evaluation makes use of an external evaluator (someone who is not within the project or initiative)
642	both qualitative and quantitative measures of outcomes of policy implementation
643	Developmental evaluation approaches should be explored as a possibility with all new policy initiatives
644	that while RCT methods have a place in evaluative activity, they must not be endorsed as 'the' way of undertaking evaluation
645	Multiple methods should be valorised
646	recommendations for translation of research to practice.
647	inclusion of logic models and theories of change.
648	The expectation to work with and gain insight from all stakeholders, from funder to consumers.
649	cultural diversity produces 'different ways of knowing' - these different ways of knowing should be incorporated at the highest levels of evaluation policy
650	a policy for the amount of money which should be allocated to evaluation in each project
651	recommendations about criteria for evaluators. Something that identifies those knowledge and skill sets that are primary requirements for someone to be an evaluator.
652	Evaluation should be linked to the original project goals. Modifications to those goals must be made "in writing" and be explicit.
653	the level of rigor in evaluation methods is appropriate to the type of program..., meat and potatoes being ramped up, done in a different location, or simply being done (vaccination program, etc) not be evaluated with a double blind randomize control trial. Only new interventions not proven be subjected to that level of rigor.
654	The list of Standards is so comprehensive that I cannot readily another policy to be added.
655	Process that will allow the program directors to formulate a formal response to the results of evaluation.
656	that only data useful for the community and for the program be collected.
657	standards of practice

658	recognition that there are a variety of methods that are appropriate to evaluation
659	making evaluation a reasonable and manageable expectation
660	a rights-based approach to evaluation, where each participant is given dignity and value
661	allow for flexibility of design
662	human rights
663	to have the evaluation appropriate to the intervention and to think about issues of full implementation which may face may issues of model fidellity in the real word - in other words is the evaluation robust enough to talk about real world implementation
664	diversity awareness. Whenever possible, an evaluation should include discussion of the minority composition of the group studied. The evaluators should report whether the general findings apply also to minority groups.
665	Evaluation efforts should include both on-going formative evaluation and summative evaluation.
666	Evaluations should contribute to learning in the field as well as knowledge about the effectiveness of a model, program or intervention.
667	Evaluations should be designed so that they can accumulate information across multiple efforts.
668	a policy outlining the goals for each evaluation study
669	a policy that outlines that all funded programs be evaluated, including a set of operational definition of terms
670	Why should there be comprehensive federal evaluation policies? I have never thought about that before. For example, should the education and public health sectors have one comprehensive set of evaluation policies? If its "policies" then it would conceivably have more "teeth" and be "required," than if it were "useful guidelines." It would be difficult to (a) develop and then (b) enforce any sort of comprehensive set of "policies" - read as "requirements" - between agencies as different as education, public health, forest resources, FDA. I"m not immediately sure it would be useful, practical to develop and esp. to implement, on first blush. I teach evaluation in a school of public health, but am not convinced this would be helpful or useful.
671	definition of a qualified evaluator.
672	Those designing the evaluation should be able to recognize and articulate the frame of reference/body of theory that informs the design (and thereby inevitably colors the results)
673	Involvement of end-users when applicable
674	Resources needed (infrastructure, financial, other) must be clearly specified and justified prior to beginning an evaluation
675	Deliverables (reports, data sets, policy recommendations etc.) must be clearly defined prior to beginning an evaluation.
676	Evaluations must be cost effective/efficient
677	Assessments and Evaluation Methodologies must accurately correspond to an explicated logic model/program theory

678	a detailed scope of work that clearly sets out the nature of the program to be evaluated and key questions to be addressed by the evaluation.
679	Evaluation should be conducted by experienced, unbiased evaluators using scientifically valid methods and objective data.
680	Explicated Logic Model/Program Theory
681	methods must be clearly delineated; likewise for data sources which should be available to others for replications of the original work
682	the involvement of independent evaluators from the outset through to completion.
683	a clear statement of what the evaluation question was and who the client is
684	transparency in evaluation methods and results
685	how results will be used to improve the program (formative evaluation)
686	be flexible with regard to methodology
687	Evaluation should identify and focus appropriate research methodology on the top priority outcomes of a program or policy.
688	the names of the authors of the evaluation should be available, even if they aren't published on the evaluation
689	information on how the results will be made public or will be classified
690	Minimum requirements on what an evaluation plan should include, suggested methodology for specific types of programs or grant types
691	reasonable % of budget designated for evaluation
692	orientation to user needs and development issues
693	ethical considerations
694	a logic model requirement for all programs
695	consistent methodology
696	qualitative methods should share equal footing with quantitative methods
697	scientific rigor is important but so is methodological flexibility
698	evaluations should be grounded in social science methodology
699	instead of just using set percentages to fund evaluation for grants for applied areas like 21st CCLC, there should be step-wise increments related to the complexity of the eval plan. So say 5-8% just to monitor the information and do only grant required evals, and 9-15% for plans with additional evals like surveys or large scale focus groups.
700	Focus on methodologies
701	Explicit qualifications for evaluators.
704	that the goal of the evaluation is to serve the citizens or the members
705	A stakeholders advisory committee with multiple perspectives
706	Provide programs that are beign evaluated a chance to get somethign that inmprove their workout of the study
707	reasonable funding amounts earmarked for evaluation
708	having professional evaluators involved in reviewing the evaluation plans submitted on federal grants
709	provisions for program evaluation and compensation at adequate levels

710	protections for the independence and objectivity of evaluators (e.g., guidance regarding agency review of evaluation reports and presentation of findings)
711	inclusion of service recipients in needs assessments
712	compliance with the Guiding Principles for Evaluators
713	consultation with expert (or at least trained) evaluators on the design, conduct,analysis, and reporting of evaluations
714	awareness that the type of evaluation method and the evaluation questions should match the stage of the program
715	a move away from RCTs as the only viable way to conduct research and evaluation.
716	the inclusion of funding for internal and external evaluation activities
717	a focus on utilization
718	A requirement for regularly scheduled evaluation of program effectiveness, and a report describing the outcomes.
719	evaluations should conform with accepted professional principles for rigor, utility, and ethics
720	emphasis on rigously evaluating efficacy of programs before implementation, and effectiveness during and after implementation.
721	the use of the National Academies (or at least their methods) to evaluate government R&D programs
722	evaluators are well versed in methodology for both naturalistic and experimental paradigms
723	special methods to evaluate high-risk, high-failure-rate, high-reward government activities
724	all projects representing investment of federal funds over a period of 3 or more years should be evaluated every 3-5 years
725	recommended use of expert judgment methods for R&D program evaluation
726	continued emphasis on evaluation of the relevance, quality, and impact of R&D programs
727	a separate set of policies for R&D operations
728	one that recognizes that not only are values and criteria culturally-informed (if not -embedded), but our ways of knowing are also culturally-embedded, and the best way to account for the privilege afforded to the scientific method, etc., especially relative to the cultures of most program participants, involves ensuring that evaluations make use of multiple members on a team, multiple disciplines, and multiple methods.
729	clear distinctions between (a) drawing inferences from program evaluation information about achievement of agency missions and goals, and (b) how such information may be validly and responsibly used in the performance appraisals of executives and employees.
730	a mandated level of independence of the reviewers and the review process
731	Evaluation methodology
732	wariness of giving OMB too much influence over specific evaluations and evaluation criteria, because OMB is "responsively competent" to the policy preferences of a President, which are usually biased in one direction or another.

733	Evaluation criteria and standardization
734	Respect of any disabilities that the participants may have.
735	Respect of diversity and cultural background of participants
736	the importance of confidentiality
737	recognition that not all evaluations will facilitate programmatic decisions. Evaluations are certainly helpful for decisions, but results are unpredictable. Frequently, you need results from several evaluations in order to put together a good picture of what is going on, and how you might improve. Therefore, it is important that additional purposes be explicitly acknowledged, including (1) learning and (2) oversight. Evaluations help inform inform Congress, the President, political appointees, career executives, and nonfederal stakeholders in their thinking, roles in policy making, and roles in oversight of federal policies.
738	the importance of ethics
739	Ownership and inclusion of the participants of the evaluation and those who are studied
740	An evaluation of the federal program itself must be undertaken at periodic intervals, not just of grantee programs or interventions.
741	Upfront funding of evaluation in federal legislation but not just jobbed out to huge evaluation firms for mega evaluations. Act globally, fund locally
742	focus on public policy objectives, for sure, but also management capacities and performance in agencies. Good processes influence end outcomes.
743	greater recognition that performance measurement frequently is driven by factors other than a program or intervention, and that "performance measurement" is in many ways a misnomer, because only some of the "performance" has been determined by the intervention.
744	a five-year cyclical evaluation of the outcomes from federal evaluation policies.
745	the explicit addition to the scope of evaluation requirements tools such as tax expenditures, voluntary regulation (e.g., by an industry), regulations, etc., that may not have budget sums associated with them.
746	some separation of evaluation policy from being subsumed primarily in the budget process. A "snapshot" of current knowledge about evaluation information certainly should be used at various stages of the budget process (e.g., bureau requests to the department, departmental requests to OMB, presidential requests to Congress), but evaluation is something that should be used continuously during the planning and management of federal government policies and activities, not just during certain "time ghettos" of the fiscal year.
747	Respect of diversity of methodology
748	engage with Congress and political appointees about the importance of not arbitrarily cutting back on evaluation capacity. It's like giving your own government a lobotomy.
749	require evaluations to explicitly acknowledge the major ways in which major stakeholders have defined "success", in order to place results of an evaluation in this broader political perspective. I'm using the term "political" in a positive, not pejorative, sense.
750	guidelines for engaging evaluators including the compentencies required
751	Require external evaluators in order to maintain objectivity

752	serious rethinking of the need for "independence" in evaluation. Many internal evaluations are necessary for good program management and strategic thinking, and should not necessarily be outsourced within an agency or to outside entities. Provide a check on internal evaluations by subjecting them to oversight. I wonder how much of the drive toward independence is fed by financial interests of some groups of actors.
753	consultation with a trained evaluator
754	the importance of developing an evaluation methodology appropriate for the program, the purpose of the evaluation, the populations involved and the resources available.
755	all publicly funded programs, etc. to include a formative (implementation) and summative evaluation component which is publicly reported upon completion of the evaluation.
756	assess the adequacy of the federal evaluation workforce (albeit with the risk that in doing so and identifying the workforce, some political actors tend to see evaluation as "administrative expenses" that should be seriously cut).
757	provide clarity in statute, OMB guidance, and agency guidance about intended audiences of evaluations and performance measures. Evaluations and metrics cannot be all things to all people, or else agency personnel get whip-sawed trying to please everyone (and no one).
758	Good faith involvement of an array of stakeholders
759	The assignment of X% of the budget for evaluation activities.
760	Organizations must be required to hire external evaluators
761	require agencies to consult with Congress and nonfederal stakeholders on evaluation agendas and results. Also require most evaluations to be posted on the Internet (perhaps with the exception of some management-oriented evaluations).
762	evaluation standards consistent with the Joint Standards
763	continuing education for Congress, political appointees, and career executives on the distinction between "program evaluation" and "performance measurement."
764	risk-based prioritization of evaluation coverage (given that we do not have enough resources to evaluate everything).
765	establish agency "chief program evaluation officers" (CPEOs), with certain qualification requirements, including not being selected on the basis of partisan affiliation, and including being career employees (and perhaps setting a term length, to prevent people from being in the positions too long). And establish a council of the officials with reporting requirements to Congress and the President.
766	that all evaluators must be neutral to eliminate bias in any resulting data to the stakeholders.
767	statutorily require evaluations to be conducted without political influence (like the National Assessment of Vocational Education: Final Report to Congress, June 2004).
768	a mandatory investigation of fidelity to implementation, especially when doing summative evaluation
769	to include end users in the evaluation planning process for program evaluations

770	Continued requirement that federal grantees include summative/formative evaluation plan as part of applications and, if funded, operations; feds allow for flexibility in evaluation measures and methodologies, especially in demonstration projects; encouraging use of third-party evaluators in evaluation design and implementation activities; and that the feds make use of the results from these evaluations.
771	redefine the "gold standard" for study design from RCT to fit between context and methods
772	a mandate that the data used for the evaluation be made avaiable to the public once the evaluation in published.
773	emphasize the growing recognition of the importance of mixed methods.
774	a requirement that most evaluations should use a random experimental design.
775	Congress establishing criteria in law somewhat like Canada has done (http://www.oag-bvg.gc.ca/internet/English/meth_gde_e_10217.html), though in less detail, which would address many criticism of the Bush Administration's Program Assessment Rating Tool (PART). OMB would then need to establish guidance documents for agencies.
776	disclosure of political activities, lobbying, and funding from evaluation companies/individuals.
777	the budget for proposed projects should be required to reflect a set percentage for evaluation
778	Congress establishing a "statutory office" in the Office of Management and Budget (OMB) that is focused on evaluation policy, and providing for an Administrator to head the office. It would be helpful if Congress legislated in some detail by providing some description of what is "in scope" in evaluation policy (e.g., to prevent advocates of one method to dominate other methods) and made the position career civil service, instead of politically appointed (to prevent politicization), and subject to certain qualification requirements. This is wortwhile, because OMB has little institutional capacity to conduct, direct, or interpret program evaluations, as exemplified by how they early-on adopted RCTs as the most "rigorous" method and did not reflect the literature on how the use of different methods depends on the underlying evaluation question.
779	compliance with set of ethics for the evaluation profession
780	to the extent possible, develop common evaluation measures for similar programs. This has already been done to some extent with the GPRA measures.
781	disclosure of both statistically significant and practically significant findings, when applicable.
782	more emphasis on transparency of the limitations of any particular methodology and its implications within reporting of results and recommendations
783	the cost-effectiveness of the "thing, phenomenon" evaluated.
784	evaluation options are tailored to the project- no "absolute" or "gold standard" is pushed one way or the other.
785	including evaluation purposes and plans in proposals. flexibility in methodological choices.

786	requiring federal agencies to publish draft versions of certain things required by the Government Performance and Results Act of 1993 in the Federal Register for public viewing and comment, and requiring agencies to comment on what they receive, in order to provide a transparency check against politicization of agency goals and metrics and allow for better congressional oversight. Topics to be subject to this requirement would be agency (1) strategic plans (including mission statements, goals, program evaluation agendas, etc.) and (2) performance goals and performance indicators to be included in annual performance plans. This is a worthwhile thing to do, because definitions of "success" are politically contentious, especially in the USA's "separation of powers" system, and because evaluations are best conducted when there is some consensus on the appropriate goals of public policies.
787	that organizations receiving government funding have a mandatory evaluation plan in the funding proposal with expected costs.
788	inclusion and meaningful participation of local evaluators in the design of grant-funded program evaluations.
789	make the evaluation framework as similar as possible across agencies -- at least to the extent that the same kind of information must be tracked
790	guidelines for when external evaluation is necessary.
791	a required measure of quality for projects that are completed under budget to determine if efficacy was sacrificed for financial savings.
792	additional evaluation funds are provided outside of program budget costs.
793	focus on outputs to the extent possible
794	that every program be evaluated
795	consultation with a trained evaluator.
796	relevant, practical, and low cost options are considered before costly and impractical evaluation options are sought.
797	time limits, e.g. evaluation of social programs much happen before XX time goes by.
798	evaluation starting from project inception.
799	Privacy: the specific level(s) of privacy in each phase of the study should be planned by evaluation team & evaluee. where on the continuum of full disclosure....-> confidential....-> anonymous; should be purposeful and explicit.
800	protections for vulnerable populations if involved in the evaluation.
801	integration of an evaluation component during any program development processes.
802	a plan for when (formative and summative) and how evaluations should be conducted
803	OMB PRA exemption for collecting information from anyone receiving funding through the Federal program being evaluated.
804	a mixed-method approach to evaluation questions with supporting quantitative and qualitative techniques
805	evaluation team should have both external and internal (to the program being evaluated) membership. should be seen as a collaborative effort.
806	required evaluator experience in the field to be evaluated.

807	Emphasis on utilization and value added. Evaluation is a tough sell under good circumstances, a history or perception that the evaluation is for academic purposes dramatically reduces the quality and efficacy of the evaluation.
808	decide on how the results will be used BEFORE the study. communication and use of results should be part of the initial planning.
809	a focus on evidence-based policy making and program implementation
810	understand that evaluation is an intervention, that is, any way one observes a phenomenon also changes the phenomenon. try to make the change in a positive direction.
811	the importance of triangulation; using multiple methods to draw valid conclusions/
812	a "so what" section in reports that transfers findings, theoretical or otherwise, into practice.
813	privacy
814	use many different ways to examine a given question: quantitative, qualitative, broad/narrow, shallow/deep. convergent validity.
815	an assessment-based strategy to evaluations: objectives linked to goals, measures linked to objectives, findings linked to measures, achievement targets linked to findings, and so on.
816	issues of equity and diversity in the programs being evaluated.
817	a focus on evidence-based evaluation, divorcing it from evaluations driven by politics or "expected outcomes" of programs
818	if evaluations are to be required, the funds to conduct them must be provided
819	evaluator or evaluation company's primary place of business must be based in the United States.
820	situation-specific quantitative and/or qualitative methodology for the evaluation at hand, rather than subscribing to a cookie-cutter method, with a greater focus on process and formative evaluation over summative evaluation.
821	the evaluation results must be considered in future policies
822	an evaluation team should include relevant and appropriate expertise
823	the evaluation measure the spesific mission statetment,the goals and the outcomes of the projec
824	federally funded evaluations should be made readily available to the research community for study and metaevaluation.
825	increased emphasis on evaluation for program improvement rather than only for funding decisions.
826	easily navigable clearinghouse of evaluation reports (data sharing) that is openly accessible to anyone interested
827	complete disclosure of evaluation findings and costs associated with publicly funded projects.
828	inclusion of evaluation experts in formulating major evaluation-related statements, positions, policies, and programs.

829	a centralized respository or tool that digitally archives all evaluation reports/products/findings producted as part of Federal projects or programs. This repository should house all of the products that are officially submitted as part of a given project/program.
830	flexibility to choose the methods and research design that fit the program and context
831	guidelines regarding the level of investment (cost) that might be required for various levels of evaluation rigor/methods. Currently, most evaluation embedded within program budgets is underfunded.
832	technical assistance to foster evaluation capacity-building within orgs/agencies
833	a "better business" or "consumer protection" listing/rating with feedback capability on all evaluation companies, groups, and individuals working with publicly funded evaluations.
834	that methods should be appropriate to the evaluation questions at hand, with no "gold standard" method.
835	Evaluation is supported as more than a means for compliance, but can be useful to grantee organization/agency decision making
836	For grant funded activities, the evaluation requirements of a given program/project should be in realistic proportion for the scale/scope of the project.
837	adherence to a set of ethical principles
838	suggested or recommended evaluation approaches or measures for types of programs, but not required
839	an independant agency, unbias and influenced by policy and funding to insure unbias and objective mandate, that priortizes and systematically addressed attainment of pre-established goals and benchmarks of other agencies
840	culturally competent data collection methodologies
841	that internal and external (i.e. those done by organizations independent from the funders) evaluations can be appropriate, depending on the context of the evaluation
842	evaluation is not the 'end-all' solution to informing policy decision-making, but rather is one integral source of the various types of information relied upon.
843	utilization of mixed-method approaches.
844	sufficient funding to support sufficient rigor (quantitative) and trustworthiness (qualitative)
845	Attend to tribal-based American Indians
846	the size of the evaluation should correspond to the size of the program being evaluated
847	support for evaluation capacity building within organizations.
848	Attend to urban-based American Indians
849	a recognition of evaluation as a fundemental element of program planning and administration, not a stand-alone component. As such, evaluation should be holistically considered in the design of a given program or project.
850	racial/ethnic categories (avoid "color-blind" policies
851	that evaluations should be linked to program goals and logic
852	support and encouragement for using a systems approach to evaluation.
853	methods serve evaluation questions

854	support for organizations that do not currently use evaluation to adopt it.
855	that an evaluation plan should include a definition of the intended use of the evaluation results, e.g. for program or site improvement, for general learning about "what works", for future funding decisions, for evaluation of a larger portfolio of projects
856	incentives for systematically meshing evaluation planning with program planning.
857	a reduction in paperwork and policies.
858	that attention is given to racial disparities in outcome measures.
859	data sharing policies that make it easier for researchers to access data related to federally-funded projects.
860	expectations for various standards and levels of evaluations.
861	specific plans for evaluation use.
862	The provision of evaluation set-aside funds that agency offices can appply to use to support internal evaluations
863	a set of common measures and format for evaluations as well as external evaluation requirements and funding for quality evaluations.
864	the evaluation measure the specific goals and outcomes of the project
865	that the evaluation be relevant both to the program and the underlying policies that are addressed.
866	formative and summative evaluation
867	qualitative and quantitative data collection activities
868	mandatory external evaluation requirements
869	explicitly allowing of a variety of evaluation methods (e.g. not requiring a "gold standard" in cases where it is not feasible or not required)
870	components of evaluation report
871	clear disclosure of brainstorming and alternative methods considered
872	format of evaluation report
873	the encouragement of emerging culturally specific agencies to participate in research and evaluation study.
874	payment range for evaluation activities
875	selection/creation of an evaluation design that matches the evaluation question to be answered
876	Pre/post/follow-up data collection should not be the de facto evaluation design. Evaluation designs should be as rigorous as possible, given whatever political, financial, and practical constraints exist in a particular situation.
877	that each program has formulated a plan for use of the results
878	evaluations should be conducted or verified by independent people/organizations.
879	an appropriate level of human subject protection, neither ignoring considerations of risk, nor holding straightforward evaluation activities to standards more appropriate for risky research.
880	Evaluations should be multidisciplinary, and not just the domain of positivists, and mixed methods should be the standard process for understanding value.
881	outsourcing the evaluation to a private agency not reliant on the agency for income.

882	Process evaluation-knowing the effects of the factors in the "black box."
883	that program clients be included as stakeholders whenever possible
884	a template for what constitutes a "good" or "comprehensive" evaluation
885	Guidelines for human subject protection/IRB requirements that make sense for low-risk evaluation studies.
886	the evaluation questions should come directly from the funded proposal and then those questions drive the evaluation design and methodology (not a standard experimental design expectation)
887	multiple methods should be encouraged, particularly combining qualitative and quantitative approaches
888	The notion of Generalizability should be transformed into the recognition of contextual and significant local realities, rather than seen as a form of Truth for national policy.
889	emphasis should be placed on both process and outcome evaluations.
890	the methods should be appropriate to a programs environment and stage of development
891	a policy about what constitutes a "good" (i.e., methodologically sound, useful, ethical) evaluation
892	guidelines for the appropriate use of evaluation designs depending on the life stage of the program (developmental, formative, summative, implementation fidelity, etc) and the information needs of the stakeholders and primary intended users.
893	a common framework that establishes guiding principles yet is flexible enough to accommodate unique program charactertistics.
894	focus on use of evaluation data/information for local program improvement, for federal grant program improvement and for informing congress and other stakeholders
895	outcomes should be a primary focus of evaluation plans
896	that a standard core set of evaluation policies should drive all federal evaluations
897	A set of common measures
898	principles of participatory action research should be included in evaluation design
899	federal reporting mechanisms should be in place that allow for information entry that makes sense with a variety of data collection and analysis methods and that allows for good reporting back to congress. I am thinking of the ED 529 form which doesn't work well with quasi-experimental evaluation studies.
900	adding a mechanism to see how/if findings from evaluations are used.
901	evaluations should be responsive to "subjects" of projects being evaluated not just project staff
902	Equity and diversity in the grants asignation
903	acknowledging the difference between research-focused studies and the evaluation-focused studies. In the former, the goal is to find a truth or prove the "effectiveness" of the treatment group over the control. In the latter, the goal is to provide useful information to the primary intended users.
904	Organizations should create logic models to guide their evaluations.
905	that a toolbox or clearinghouse of tested evaluation methods and techniques be made readily available to programs required to complete evaluations.

906	evaluators need to understand that every evaluation is a political event
907	that evaluation plans should be reviewed by evaluation professionals for completeness and appropriateness to the program being evaluated.
908	that programs or organizations receiving federal funding should be required to submit an evaluation plan that is appropriate for the level of the project's development.
909	setting common measures across types of programs.
910	encourage the use of mixed method designs
911	the development of an evaluation plan which describes how each federal program intends--and does--evaluate itself
912	appropriate funding of evaluation as a part of funded programs
913	mandatory evaluation for funded programs.
914	a emphasis on high-quality, methodologically rigorous (and diverse)evaluation that informs policy making. Such evaluations should emphasize the judgmental nature of evaluation (e.g., good is good, bad is bad, and it's the evaluators job to determine which is which) rather than mere factual (research-based) statements and premises.
915	room for methodological diversity.
916	there should be a mechanism by which federally funded programs track evaluation systems - with the goal of developing a set of evaluation systems that - when applied appropriately to similar programs in different settings - comparisons can be made.
917	that any federally funded program should have a system in place for feedback and improvement.
918	...that all agencies or organizations receiving government funding have a mandatory evaluation plan in the funding proposal.
919	that any federally funded program should be evaluated appropriately for its level of development
920	value should be placed on informal evaluation, open-ended responses, and non-traditional methods of evaluation.

Appendix B-List of compound statements with break-out statements

ID#	Compound statement		Break-out statement
101	Evaluation designs & methods should allow for different levels of complexities in change theories (simple, complicated, complex, chaotic; see Cynefin Framework)	101a	Evaluation design & methods should suit the context of the evaluation
		101b	Evaluation designs & methods should allow for different levels of complexities in change theories (simple, complicated, complex, chaotic; see Cynefin Framework)
120	That the needs of the evaluation and the questions that are asked drive the methods, and that all methods are equally valued according to the purpose of the evaluation	120a	That the needs of the evaluation and the questions that are asked drive the methods
		120b	that all methods are equally valued according to the purpose of the evaluation
132	evaluations can be conducted both externally and internally, but both must include conditions of independence and non-retaliation. Separate contracts should be included with major proposals to ensure that funds available for evaluation are not reallocated to program efforts prior to completion.	132a	evaluations can be conducted both externally and internally, but both must include conditions of independence and non-retaliation
		132b	Separate contracts should be included with major proposals to ensure that funds available for evaluation are not reallocated to program efforts prior to completion.
13	External evaluation is no more valid than "internal;" the rigor by which the design and implementation of the evaluation is done determines its validity.	13a	the rigor by which the design and implementation of the evaluation is done determines its validity.
		13b	External evaluation is no more valid than "internal;"

156	1. Develop evaluation competencies that would define professional preparation skills, standardize evaluation language, and develop competency-instruction workforce education and training. 2. Requirement for federally funded programs to include evaluation. 3. Evaluation data should feed back into program management. 4. Attention to mixed methods; respect for qualitative, quantitative, and participatory evaluation. 5. Attention to evaluation use - how results will be used for program improvement.	156a	Attention to mixed methods; respect for qualitative, quantitative, and participatory evaluation.
		156b	Requirement for federally funded programs to include evaluation
		156c	Evaluation data should feed back into program management.
		156d	Attention to evaluation use - how results will be used for program improvement.
		156e	Develop evaluation competencies that would standardize evaluation language
		156f	Develop evaluation competencies that would develop competency-instruction workforce education and training
		156g	Develop evaluation competencies that would define professional preparation skills.
158	at the macro federal level, there needs to be a movement to standardize the evaluation language. So that the evaluation work funded from one federal agency can be compared with the work from and other agency. Also, when and if there are professional-wide adopted professional competencies for evaluators with some degree of common preparation standards, there could be some possibility of being recognition as a "professional" class by the feds.	158a	at the macro federal level, there needs to be a movement to standardize the evaluation language. So that the evaluation work funded from one federal agency can be compared with the work from and other agency. Also, when and if there are professional-wide adopted professional competencies for evaluators with some degree of common preparation standards, there could be some possibility of being recognition as a "professional" class by the feds.
		158b	when and if there are professional-wide adopted professional competencies for evaluators with some degree of common preparation standards, there could be some possibility of being recognition as a "professional" class by the feds

187	appropriate resources be allocated to ensure evaluation at all stages of program	187a	appropriate resources be allocated to ensure evaluation at all stages of program
		187b	appropriate funding to match the importance of the policy or the commitment of taxpayer resources.
208	I agree with culturally competent and utilization-focused evaluation that is disseminated for implementation of results	208a	I agree with culturally competent evaluation
		208b	utilization-focused evaluation that is disseminated for implementation of results
229	budgeting for robust evaluation that is utilization-focused and culturally competent.	229a	budgeting for robust evaluation
		229b	evaluation that is culturally competent
		229c	evaluation that is utlization-focused
22	the use of trained, independent evaluators who are provided with sufficient resources to execute comprehensive, multi-method evaluation designs.	22a	to execute comprehensive evaluation designs
		22b	the use of evaluators who are provided with sufficient resources
		22c	to execute multi-method evaluation designs.
		22d	the use of trained, independent evaluators
23	use of multiple data sources or methods in conducting evaluations of student learning outcomes	23a	use of multiple data sources or methods in conducting evaluations of student learning outcomes
		23b	use of multiple methods in conducting evaluations of student learning outcomes
248	I agree with a lot of statements below, such as "guidelines regarding the level of investment (cost) that might be required for various levels of evaluation rigor/methods. Currently, most evaluation embedded within program budgets is underfunded." or An accreditation process and licensing requirement for evaluators. or Requirement for Transparency in the evaluation process or Requirement of participation by end-users	248a	Requirement for Transparency in the evaluation process
		248b	Requirement of participation by end-users
		248c	An accreditation process and licensing requirement for evaluators.
		249a	Ethical treatment of individuals

249	Qualifications of evaluators Ethical treatment of individuals Adherence to some set of professional evaluation standards	249b	Qualifications of evaluators Ethical treatment of individuals Adherence to some set of professional evaluation standards
		249c	Adherence to some set of professional evaluation standards
271	the design of the evaluation should be appropriate to the research questions, employing methodologies that are rigorous and are a good fit (ethically and methodologically) with the program content and context.	271a	employ methodologies that are a good fit with program duration, intensity, lifecycle stage, context, available capacity, stakeholder needs,
		271b	the design of the evaluation should be appropriate to the research questions, employing methodologies that are rigorous and are a good fit (ethically and methodologically) with the program content and context.
277	valuing a wide range of methods (qualitative & quantitative) and research questions (implementation & impact).	277a	valuing a wide range of methods (qualitative & quantitative) and research questions (implementation & impact).
		277b	valuing a wide range of research questions (implementation & impact).
292	Qualified persons to design and conduct an evaluation that adheres to AEA guiding principles, is appropriate for the project/location/and questions to be answered, and follows a model generally accepted in the program evaluation community. Adequate funding should also be earmarked specifically for evaluation activities.	292a	Adequate funding should also be earmarked specifically for evaluation activities.
		292b	evaluation that adheres to AEA guiding principles
		292c	follows a model generally accepted in the program evaluation community
		292d	appropriate for the project/location/and questions to be answered
		292e	Qualified persons to design and conduct an evaluation.
29	evaluation questions and program goals that are: 1) useful to the funder, the program, AND the community served, and 2) realistic given the duration, intensity, and context of the program	29a	program goals that are realistic given the duration, intensity, and context of the program

		29b	evaluation questions that are realistic given the context of the program
		29c	evaluation questions and program goals that are useful to the funder, the program, AND the community served
31	evaluation should be a standard part of program development and implementation, included from the very beginning of program selection and design... from initial formative evaluation through outcome and impact evaluations	31a	evaluation should be included from the very beginning of program selection and design
		31b	evaluation should be a standard part of program implementation
329	The decision-makers and types of decisions to be supported by the evaluation should be clear. The methods should be appropriate to the purpose, resources and intended completion date of the evaluation.	329a	The methods should be appropriate to the purpose, resources and intended completion date of the evaluation.
		329b	The decision-makers and types of decisions to be supported by the evaluation should be clear.
330	should have a minimum budget allocation for programs. Also restructure the way they do RFPs - they typically solicit evaluators to conduct methodology exercises (surveys, focus groups, etc.) rather than focus on a specific (and useful) research questions and indicators.	330a	should have a minimum budget allocation for programs. Also restructure the way they do RFPs - they typically solicit evaluators to conduct methodology exercises (surveys, focus groups, etc.) rather than focus on a specific (and useful) research questions and indicators.
		330b	restructure the way they do RFPs - they typically solicit evaluators to conduct methodology exercises (surveys, focus groups, etc.) rather than focus on a specific (and useful) research questions and indicators.
35	context and implementation is equally as important as outcomes	35a	context is equally as important as outcomes
		35b	implementation is equally as important as outcomes
369	a requirement that all new programs be evaluated from the time of establishment; a requirement that all existing programs be periodically evaluated for outcomes and impact within a specified period of time not to exceed 10 years.	369a	a requirement that all new programs be evaluated from the time of establishment; a requirement that all existing programs be periodically evaluated for outcomes and impact within a specified period of time not to exceed 10 years.

		369b	a requirement that all existing programs be periodically evaluated for outcomes and impact within a specified period of time not to exceed 10 years.
370	a specific plan for disseminating the results to key program stakeholders and how the results will be used. Furthermore, those stakeholders should be identified at the beginning of the evaluation.	370a	a specific plan for disseminating the results to key program stakeholders and how the results will be used. Furthermore, those stakeholders should be identified at the beginning of the evaluation.
		370b	stakeholders should be identified at the beginning of the evaluation.
394	information to contextualize findings should be collected, and results should be disaggregated to inform policy in different settings.	394a	information to contextualize findings should be collected, and results should be disaggregated to inform policy in different settings.
		394b	results should be disaggregated to inform policy in different settings.
405	to ensure that the evaluation is funded and external to the agency conducting the program.	405a	to ensure that the evaluation is funded
		405b	to ensure that the evaluation is external to the agency conducting the program.
425	ensure reasonable funding for evaluation and score proposals based on sufficient funding alloted to evaluation	425a	ensure reasonable funding for evaluation
		425b	score proposals based on sufficient funding alloted to evaluation
45	that evaluation rigor/design/methods should reflect the evaluation goals/uses, scope of the project/program, and available respurces/expertise.	45a	that evaluation rigor/design/methods should reflect available resources/expertise
		45b	that evaluation rigor/design/methods should reflect the evaluation goals/uses, scope of the project/program
488	the requirement to work at the community level and any agency being funded need to meet minimum standards in order to carry out evaluation. There needs to be clear goals and an infrastructure to put in place so that agencies are accountable for what they say they will be doing.	488a	the requirement to work at the community level
		488b	any agency being funded need to meet minimum standards in order to carry out evaluation.

		488c	clear goals and an infrastructure to put in place so that agencies are accountable for what they say they will be doing
489	an evalaution framework that includes stakeholder input with clear golas and accountability measures	489a	clear goals and accountability measures
		489b	an evaluation framework that includes stakeholder input
506	a requirement for both external AND internal evaluation as well as the dissemination of evaluation findings to the general public and other programs.	506a	dissemination of evaluation findings to the general public and other programs.
		506b	a requirement for both external AND internal evaluation
524	adapting the standards to resources available in the communities or to methods acceptable in the populations.	524a	adapting the standards to resources available in the communities
		524b	adapting the standards to methods acceptable in the populations
536	programs have targets and are held accountable; data is transparent; evaluations are conducted and results are used for quality improvement	536a	data is transparent
		536b	programs have targets and are held accountable
		536c	evaluations are conducted and results are used for quality improvement
537	a standard framework for evaluation such as CDC's framework Judgment of the merit or worth of a program, policy, strategy, event, or initiative Whenever appropriate, evaluators should include key stakeholders as contributors to the evaluation plan and implementation	537a	a standard framework for evaluation such as CDC's framework
		537b	Judgment of the merit or worth of a program, policy, strategy, event, or initiative
		537c	Whenever appropriate, evaluators should include key stakeholders as contributors to the evaluation plan and implementation
539	the need for integration of the Program Evaluation Standards. A firewall should be established between the evaluators and those who have a vested interets in the outcome of the evaluation.	539a	A firewall should be established between the evaluators and those who have a vested interets in the outcome of the evaluation.

		539b	the need for integration of the Program Evaluation Standards
53	A commitment to building evaluation capacity and an evaluation culture	53a	A commitment to building evaluation capacity
		53b	A commitment to building an evaluation culture.
586	Procedures for evaluation; transparency and fairness of evaluation	586a	Procedures for evaluation
		586b	transparency and fairness of evaluation
590	evaluation should be included from the design phase and budgeted for appropriately	590a	evaluation should be included from the design phase and budgeted for appropriately
		590b	evaluation should be budgeted for appropriately
591	that every federal program should consult with a team of experts in evaluation and create a plan for how evaluations of grant initiatives will be used to inform decision making	591a	that every federal program should create a plan for how evaluations of grant initiatives will be used to inform decision making
		591b	that every federal program should consult with a team of experts in evaluation
609	evaluation of clearly stated outcomes included in the program design. Outcomes do not have to be only participant change. Mixed methods for determining if goals of the program were reached and clear guidance on use of the evaluation....how it will be used.	609a	evaluation of clearly stated outcomes included in the program design.
		609b	Outcomes do not have to be only participant change.
		609c	Mixed methods for determining if goals of the program were reached
		609d	clear guidance on use of the evaluation....how it will be used.

611	Build evaluation into the overall program design. Evaluation design include consideration of how the evalution will be used and who will use the evaluation results. Evaluation findings should be presented in a manner that makes then useful and accessible to stakeholders at all levels including those involved in implementing programs. cultural diversity produces 'different ways of knowing' - these different ways of knowing should be incorporated at the highest levels of evaluation policy	611a	Build evaluation into the overall program design
		611b	Evaluation findings should be presented in a manner that makes them useful and accessible to stakeholders at all levels including those involved in implementing programs
		611c	cultural diversity produces 'different ways of knowing' - these different ways of knowing should be incorporated at the highest levels of evaluation policy
613	review of process, implementation and theories underlying complex organizations policy development and decisions at micro and macro levels	613a	review of process, implementation and theories underlying complex organizations
		613b	policy development and decisions at micro and macro levels
65	(1) flexible methods/methodologies depending on the questions that drive the evaluation are acceptable if conducted rigorously, (2) minimum (~15%) requirement in all budgets for evaluations (includes formative & summative work); (3) measurement of outcomes/impacts appropriately linked to time-lines of projects (i.e., no expectation for long-term impacts for projects that have only 1 to 3 years to design, implement, and evaluate); (4) well-designed program logic models (with evaluation directly linked to these; (5) summative evaluations should document how formative evaluation information has been utilized in the project	65a	well-designed program logic models (with evaluation directly linked to these
		65b	minimum (~15%) requirement in all budgets for evaluations

		65c	(1) flexible methods/methodologies depending on the questions that drive the evaluation are acceptable if conducted rigorously
		65d	measurement of outcomes/impacts appropriately linked to time-lines of projects (i.e., no expectation for long-term impacts for projects that have only 1 to 3 years to design, implement, and evaluate)
		65e	summative evaluations should document how formative evaluation information has been utilized in the project
663	to have the evaluation appropriate to the intervention and to think about issues of full implementation which may face may issues of model fidellity in the real word - in other words is the evaluation robust enough to talk about real world implementation	663a	think about issues of full implementation which may face may issues of model fidellity in the real word - in other words is the evaluation robust enough to talk about real world implementation
		663b	to have the evaluation appropriate to the intervention
664	diversity awareness. Whenever possible, an evaluation should include discussion of the minority composition of the group studied. The evaluators should report whether the general findings apply also to minority groups.	664a	diversity awareness. Whenever possible, an evaluation should include discussion of the minority composition of the group studied. The evaluators should report whether the general findings apply also to minority groups.
		664b	Whenever possible, an evaluation should include discussion of the minority composition of the group studied. The evaluators should report whether the general findings apply also to minority groups
669	a policy that outlines that all funded programs be evaluated, including a set of operational definition of terms	669a	a policy that outlines that all funded programs be evaluated
		669b	a set of operational definition of terms
679	Evaluation should be conducted by experienced, unbiased evaluators using scientifically valid methods and objective data.	679a	using scientifically valid methods and objective data.
		679b	Evaluation should be conducted by experienced, unbiased evaluators
69	methods and questions that are useful for the funder and the grantee	69a	methods ons that are useful for the funder and the grantee
		69b	questions useful for the funder & grantee

718	A requirement for regularly scheduled evaluation of program effectiveness, and a report describing the outcomes.	718a	Require a report describing the outcomes.
		718b	A requirement for regularly scheduled evaluation of program effectiveness, and a report describing the outcomes.
761	require agencies to consult with Congress and nonfederal stakeholders on evaluation agendas and results. Also require most evaluations to be posted on the Internet (perhaps with the exception of some management-oriented evaluations).	761a	Also require most evaluations to be posted on the Internet (perhaps with the exception of some management-oriented evaluations).
		761b	require agencies to consult with Congress and nonfederal stakeholders on evaluation agendas and results
770	Continued requirement that federal grantees include summative/formative evaluation plan as part of applications and, if funded, operations; feds allow for flexibility in evaluation measures and methodologies, especially in demonstration projects; encouraging use of third-party evaluators in evaluation design and implementation activities; and that the feds make use of the results from these evaluations.	770a	feds allow for flexibility in evaluation measures and methodologies, especially in demonstration projects
		770b	Continued requirement that federal grantees include summative/formative evaluation plan as part of applications and, if funded, operations
		770c	the feds make use of the results from these evaluations
		770d	encouraging use of third-party evaluators in evaluation design and implementation activities

778	Congress establishing a "statutory office" in the Office of Management and Budget (OMB) that is focused on evaluation policy, and providing for an Administrator to head the office. It would be helpful if Congress legislated in some detail by providing some description of what is "in scope" in evaluation policy (e.g., to prevent advocates of one method to dominate other methods) and made the position career civil service, instead of politically appointed (to prevent politicization), and subject to certain qualification requirements. This is wortwhile, because OMB has little institutional capacity to conduct, direct, or interpret program evaluations, as exemplified by how they early-on adopted RCTs as the most "rigorous" method and did not reflect the literature on how the use of different methods depends on the underlying evaluation question.	778a	Congress establishing a "statutory office" in the Office of Management and Budget (OMB) that is focused on evaluation policy, and providing for an Administrator to head the office.
		778b	if Congress legislated in some detail by providing some description of what is "in scope" in evaluation policy (e.g., to prevent advocates of one method to dominate other methods)
		778c	if Congress made the position career civil service, instead of politically appointed (to prevent politicization), and subject to certain qualification requirements.
820	situation-specific quantitative and/or qualitative methodology for the evaluation at hand, rather than subscribing to a cookie-cutter method, with a greater focus on process and formative evaluation over summative evaluation.	820a	situation-specific quantitative and/or qualitative methodology for the evaluation at hand, rather than subscribing to a cookie-cutter method
		820b	a greater focus on process and formative evaluation over summative evaluation.
863	a set of common measures and format for evaluations as well as external evaluation requirements and funding for quality evaluations.	863a	funding for quality evaluations
		863b	a set of common measures and format for evaluations
		863c	external evaluation requirements

89	Evaluation happen at all levels when program funds can support the evaluation. The evaluation must focus on program development with outcome data being collected at a future date, determined to be reasonable by the program, audience and anticipated outcomes (all of which should be identified at beginning of program development). Evaluation results should be made available to peers and stakeholders.	89a	The evaluation must focus on program development with outcome data being collected at a future date, determined to be reasonable by the program, audience and anticipated outcomes (all of which should be identified at beginning of program development).
		89b	Evaluation happen at all levels when program funds can support the evaluation.
		89c	Evaluation results should be made available to peers and stakeholders.

Appendix C: Removed non-responses

1	qua	230
2	data	232
3	and fit the program and	285
4	No comment	286
5	???	287
6	If you have not already done so, see "Federal Evaluation Policy," a book written in 1974 by Wholey and others at the The Urban Institute's Program Evaluation Studies Group.	308
7	teaching students life & parenting skills and evaluating how they do on into they college years and beyond into family life.	320
8	can't think of anyting at the moment	324
9	can't think of anything at the moment	325
10	improving student achievement	331
11	a policy higher education budget cut.	378
12	clerly the following lists includes all my concerns	528
13	The list of Standards is so comprehensive that I cannot readily another policy to be added.	654

Appendix D: Final statements by super-category, with original categories

APPROACH *theoretical frameworks applied in evaluations, evaluator philosophies and biases*	
Sub-categories	Text of statements (edited form)
comprehensive (22,105,142,260,495) empowerment (515) logic model (**10**,65,144,149,472,497,545,613,627, 647,677,680,694,**904**) many methods (15,16,28,97,108,127,128,198,255,278, 288,505,505,519,548,571,588,644,645, 649,658,686,697,715,722,747,770,771, 784,869,915,920) participatory methods (18,428,620,635,673,711,739,769,**898**,901) qualitative methods (696) quantitative methods (283,334,622) reflective practitioner (316,321,**600**,672) systems approach (**314**,631,852) timeline (20,**181**)	**Ideological philosophies shall be disclosed and articulated as underlying assumptions in program theory or logic. (10)** **Require that the timing of evaluation synchronize with program design and planning rhythms. (181)** **Apply complex, systems-based evaluation methods in complex, adaptive systems. (314)** **In an evaluation, the philosophical biases of the evaluator must be clearly identified. (600)** **Evaluation designs shall include principles of participatory action research, which involves all relevant parties in actively examining together some current action in order to change and improve it. (898)** **Organizations shall create logic models to guide their evaluations. (904)**

<table>
<tr><th colspan="2">CAPACITY BUILDING
<i>training program managers and staff to do evaluation, strengthening an organization’s ability to evaluate its own programs</i></th></tr>
<tr><td>Sub-categories</td><td>Text of statements (edited form)</td></tr>
<tr><td>capacity building (53,81,<b>832</b>,847)
evaluation culture (<b>53</b>)
evaluation during development (71, 89,431,801)
incentives to evaluate (<b>259</b>)
feed forward (500)
training (<b>74</b>,262,527,529,533)
evaluator professional development (<b>78</b>)</td><td><b>Build an evaluation culture. (53)</b>
<b>Require training for federal, state and program managers--including what evaluation is, what constitutes effective evaluation work, and how to manage external evaluation. (74)</b>
<b>Support and nurture up-and-coming evaluators, through fellowships or sabbaticals at different agencies and opportunities for co-authorship. (78)</b>
<b>Encourage and value systematic internal evaluation by providing funding and bonus points to applicants with a history of focus on internal evaluation. (259)</b>
<b>Provide technical assistance to foster evaluation capacity building within organizations and agencies. (832)</b></td></tr>
</table>

<table>
<tr><th colspan="2">COORDINATION
communication and cooperation across institutional boundaries</th></tr>
<tr><td>Sub-categories</td><td>Text of statements (edited form)</td></tr>
<tr><td>clearinghouse of evaluations (166,281,282, **306**,574,826,829)
clearinghouse of methods (236, **905**)
connect and collaborate (1,4,104,119,153,165,**217**,220,221,272,298, 335,398,491,542,**563**,667,780,916,)
reassess goals (**499**)</td><td>**Value evaluation partnerships between academia and low-resource communities. (217)**
All evaluation reports conducted on federally funded evaluations shall be compiled in a central location that is open to the public. (306)
Shared understandings of program goals and mechanisms shall be re-assessed periodically. (499)
Where appropriate, agencies and programs shall use identical measures across programs and agencies. (563)
A toolbox or clearinghouse of tested evaluation methods and techniques shall be made available to programs required to complete evaluations. (905)</td></tr>
</table>

EVIDENCE:
ways of collecting evidence, ways of knowing, what constitutes truth, standards of evidence

Sub-categories	Text of statements (edited form)
cultural competence (37,42,**44**,58,84,182,208,211,227,229,242, 244,403,419,446,447,**611**,664,734,735,840, 845,848,850) mixed methods (22,23,47,61,86,110,116,118, 156,161,169,190,234,300,322,326,449,457, 474,492,561,587,609,633,642,728,773,804, 811,**814**,843,867,880,887,910) standards of evidence (**77**,**484**)	**Require that evaluations be undertaken with the cultural "lens" appropriate to support decision making. (44)** **Replace the notion of generalizability with the recognition of local contextual realities as a basis of truth for national policy. (77)** **Establish guidelines for standards of evidence. (484)** **Different ways of knowing from different cultures are valid as evidence for evaluation. (611)** **Evaluations shall use many different ways to examine a question, including quantitative and qualitative, broad and narrow, shallow and deep. (814)**

INDEPENDENCE: *coping with undue external influences on evaluators*	
Sub-categories	Text of statements (edited form)
independence of evaluation (50,90, 132,192,205,218,265,309,332,376, **504**,539,566,**710**,730,732,766,767, 817) whistleblower (**38**,454)	**Evaluators reporting evidence of coercion through the exercise of power, influence, or resources shall be protected under Federal Whistle Blowers Laws. (38)** **A contract for federal program evaluation shall be overseen by a branch or division that is over the entity being evaluated, not by the program that is the subject of the evaluation. (504)** **Establish guidelines for appropriate agency review of evaluation reports in order to protect the independence and objectivity of evaluators. (710)**

*Contracting Officer's Technical Representative

INTEGRATE EVALUATION: *connecting evaluation development with program development*	
Sub-categories	Text of statements (edited form)
evaluation during planning (31,68,91,354,**363**,368,369,410,582, 590,592,611,798,849,856) integrate evaluation (233,336,411, 568,609,619)	**Evaluators shall serve as key members of the planning body for each project and program. (363)**

META POLICY: *how to make evaluation policies, who will make evaluation policies*	
Sub-categories	Text of statements (edited form)
anti-policy (49,**215**,510,670,857,) federal evaluation personnel structure (**765**, **839**) levels of evaluation (75,89,95,488) policy levels (**55**,59,290) scope of evaluation policy (778) who makes policy (613,828) who sets standards (350,423,**708**) R&D (**727**)	**Rules and principles at the Federal level shall be devised in such a way as to set appropriate standards for evaluations at state and local levels. (55)** **There shall be no comprehensive set of evaluation policies. (215)** **Involve professional evaluators in reviewing the evaluation plans submitted on federal grants. (708)** **There shall be a separate set of policies for research and development (R&D) operations. (727)** **A "chief program evaluation officer" (CPEO) shall be appointed for each federal agency, and shall serve on a council with reporting requirements to Congress and the President. (765)** **An independent agency shall be established to prioritize and systematically address attainment of pre-established goals and benchmarks of other federal agencies. (839)**

METHOD: *how to select a method, using particular methods, who selects method*	
Sub-categories	Text of statements (edited form)
developmental evaluation (**17,643**) methods (24,64,461,679,698,700,721,723, 731,**774**)	**Developmental evaluation approaches, which allow for the redirection of any evaluation as circumstances change, shall be considered for use in all new policy initiatives. (17) (643)** **Evaluations shall be required to use a random experimental design. (774)**

<table>
<tr><th colspan="2">REPORTING:
<i>how and when to report evaluation results</i></th></tr>
<tr><th>Sub-categories</th><th>Text of statements (edited form)</th></tr>
<tr><td>integrity of findings (196)
reporting results (4,9,39,52,76,<b><u>77</u></b>,89,175,183,219,240,251,268,275,289,353,370,391,394,397,407,441,451,506,515,570,<b><u>575</u></b>,580,583,595,596,601,611,628,640,689,718,761,772,780,781,824,<b><u>827</u></b>,870,871,872,<b><u>899</u></b>)</td><td><b>Evaluation results around specific topic areas shall be published regularly by the federal government so as to identify the state of the art and any gaps in knowledge. (77)</b>
<b>Evaluation findings shall explicitly address threats to validity. (575)</b>
<b>Evaluation findings and costs associated with publicly funded projects shall be fully disclosed in a manner accessible to the public. (827)</b>
<b>Federal reporting mechanisms shall be structured in such a way as to accommodate a variety of data collection and analysis methods. (899)</b></td></tr>
</table>

RESOURCES: *funding evaluation, making strategic use of evaluation resources*	
Sub-categories	Text of statements (edited form)
benefit/cost analysis (4,40,60,112, 188,**359**,402,460,493,576,624,674, 676,783,796) evaluability assessment (88,**148**) funding for evaluation (2,**3**,4,22,34, 46,65,73,125,129,**132**,134,138,143, 144,147,171,172,173,185,187,203, 209,229,231,248,257,292,294,301, 307,317,319,330,340,342,343,356, 358,364,366,382,384,385,405,406, 416,425,455,507,531,535,572,590, 594,638,650,691,699,707,709,716, 748,759,777,792,818,**831**,844,862, 863,874,912) who pays evaluators (72)	**Grants shall include evaluation funding extending three or more years beyond the life of other grant funds, to assure follow-up. (3)** **A separate evaluation contract shall be included in major grants to ensure that funds available for evaluation are not reallocated to program efforts. (132)** **An evaluability assessment shall be conducted prior to launching into full-blown evaluation. (148)** **A cost analysis (benefit, effectiveness, utility, and/or feasibility) shall be conducted for all federal programs. (359)** **Establish guidelines for the level of funding required for evaluation conducted at varying levels of rigor, and using various methods. (831)**

SCOPE: *what should be evaluated and when things should be evaluated*	
Sub-categories	Text of statements (edited form)
what gets evaluated (4,186,**210**,246,252,253,297,320,331,378,**498**,555,609,656,687,726,**740**,745,764,793,823,858,**895**) when to evaluate (4,6,**551**,724,744,**746**)	**Evaluate to learn not just what is happening but why. (210)** **Measure all outcomes, not just expected ones. (498)** **Establish guidelines for when and under what circumstances evaluations should be undertaken. (551)** **Periodically undertake an evaluation of the federal program itself, not just its grantee programs. (740)** **Use evaluation continuously during the planning and management of federal government policies and activities, not just at budget time. (746)** **Make outcomes the primary focus of all evaluation plans. (895)**

STANDARDS: *assuring quality in the way evaluation is practiced through standardization*	
Sub-categories	Text of statements (edited form)
evidence-based (4,82,**114**,163,302,357,464,**485**,585,809) require evaluation (4,5,80,155,156,160,222,362,369,389,415,494,513,630,669,718,787,794,797,854,913,**917**) require evaluation plan (4,157,162,**274**,351,360,553,578,668,802,911,918) response rate (**543**) standard language (**158**,**315**,**328**,669) standards (4,13,41,48,63,107,115,154,224,238,**247,**254,261,264,292,292,**323,**339,341,**348,**349,365,371,375,395,413,433,440,450,471,483,486,488,509,523,**534,**537,538,539,540,541,550,**554,**556,557,558,565,586,695,712,719,733,762,775,789,863,884,891,893,**896,**897,909)	**All evaluations shall be informed by the most current knowledge about evaluation theory and practice. (114)** **Standardize the evaluation language so that the evaluation work under one federal agency can be compared with the work under another agency. (158)** **The standard for the evaluation of all social programs shall be whether the program advances the goals of social justice. (247)** **For all federally funded projects and programs, an evaluation plan must be established, documented and finalized within 60 days of funding. (274)** **Only the English language shall be used in evaluation studies. (315)** **Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare. (323)** **Data collection instruments shall be translated into languages other than English to ensure that all voices are heard. (328)** **Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy. (348)** **Explicit criteria shall be developed for identifying "promising" innovative practices. (485)** **Impact evaluations must utilize research designs that support strong causal inference. (534)** **Evaluations are required to obtain an 80% response rate. (543)** **The federal government shall establish a clear, universal working definition of the term "program evaluation". (554)** **There shall be a core set of evaluation policies which apply to all federal evaluations. (896)** **All federally funded programs must have a system in place for feedback and improvement. (917)**

TAILORING: *designing the evaluation (its approach, design, methods, measures) to fit the program and its context*	
Sub-categories	Text of statements (edited form)
consider context (35,117, 318,663) match approach to context (101,103,195,271,372,386,469,614, 621,841) match approach to feasibility (7,45, 139,417,524,659,876) match approach to goals (65,67,**86,**100,120,120,239,**271**,292, 313,329,426,430,439,470,**599**,603,652,754, 815,820,834,853,875,886) match approach to lifecycle (65,98, 111,121,174,616,714,890,892,908,919) match approach to population (263,524) match approach to program (27,29, 45,101,178,201,277,330,383,418,437, 458,526,610,653,661,690,830,836, 838,846,851,864,865) match approach to utilization (11,36, 57,62,69,133,461,611,623,692) match measure to target population (**43**)	**Evaluations shall employ measures that are reliable and valid for the respondents. (43)** **Evaluations shall employ methods and procedures that fit the focus of the evaluation. (86)** **Evaluations shall employ methodologies that fit with program duration, intensity, lifecycle stage, context, available capacity, stakeholder needs. (271)** **The level of rigor of the evaluation shall be positively related to the magnitude and criticality of the decisions to be made on the basis of results. (599)**

TRANSPARENCY: *making data or measures or evaluation methods, evaluation results transparent, making managers accountable*	
Sub-categories	Text of statements (edited form)
access to data (4,**123**,296,304,310,377, 482,803,859) transparency (4,51,179,197,248,329, 387,409,**422,**438,444,459,481,514,536, 569,586,**602,**618,625,639,681,684,688, **776,**782,**786**)	**Federally funded evaluations shall automatically provide access to data sets maintained by public entities, provided there is appropriate IRB oversight. (123)** **Scoring rubrics including evaluation plans shall be published in advance as a part of RFP application review. (422)** **Data analytic techniques employed in federal evaluations must be fully described and replicable. (602)** **Project managers must disclose political activities, lobbying, and funding from evaluation companies/individuals.** **(776)** **Require federal agencies to periodically publish draft versions of strategic plans and performance goals and indicators. (786)**

USES OF EVALUATION: *saying how evaluation results should be used and when*	
Sub-categories	Text of statements (edited form)
evaluation informs decision making (338,352,467,562,**591**) fidelity assessment (**137**) follow up evaluation (**478**) formative evaluation (30,35,65,146, 184,200,216,237,277,305,**327**,333, 380,448,461,476,501,520,632,665, 720,742,755,768,770,820,866,882,889) negative results (**12**,373) outcome evaluation (29,70) performance measurement (**8**,**26,**106, 489,**608,**743,763,791) pro-forma evaluation (**32**,79) purpose of evaluation (93,113,193,241, 279,462,479,488,496,536,537,549,607, 666,704,729,**737**,785,855,903,914) utilization (19, 21,25,150,156,156,176, 191,204,208,214,228,229,235,244,256, 266,269,284,303,337,355,361,388,392, **394**,400,408,412,421,427,442,443,452, 465,503,515,516,517,522,530,536,546, 609,629,637,646,655,685,706,717,770, 807,**808**,812,821,825,835,842,861,877, 894,900)	**Establish in advance measurable criteria by which the success of programs will be evaluated. (8)** **Value evaluation findings of "no discernable effect" as constructive feedback for program redirection. (12)** **Allow programs an appropriate period of time in operation before expecting them to achieve and report long-term, sustainable change. (26)** **Detect and decrease symbolic or pro-forma evaluations conducted solely to adhere in a minimum way to requirements for evaluation. (32)** **All federally funded programs shall conduct regular fidelity assessments. (137)** **All federally funded projects shall include both process and outcome evaluation components. (327)** **Disaggregate evaluation results to inform policy in different settings. (394)** **Every evaluation shall begin by identifying and articulating existing informal assessments and information feedback loops. (461)** **Where recommendations are enacted as a result of an evaluation, a second evaluation shall later be conducted to determine if they were successful. (478)** **Every federal program shall create a plan for how evaluations of grant initiatives will be used to inform decision making. (591)** **Review Government Performance and Results Act (GPRA) indicators at least every other year and ask the leadership of funded programs for their input. (608)** **Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight. (737)** **Every evaluation plan shall indicate how the results are to be used and communicated. (808)**

VALUES: *larger ideals that guide practice*	
Sub-categories	Text of statements (edited form)
Diversity (4) Equity (2) Ethics (26) evaluation as intervention (1) human subjects (23) minority (2)	**Evaluations shall report not only majority views and findings but also minority views and findings. (436)** **Evaluations shall be conducted with the goal of assuring that the population or program benefits from the evaluation. (525)** **Individual data gathered for one purpose may not be used for any other purpose without express permission of the individual. (604)** **Evaluation findings shall include a discussion of the minority composition of the group studied and an indication of whether the general findings apply to minority groups. (664)** **Every evaluation study involving human subjects shall specify in advance the level of privacy for each phase of the study, based on a continuum from full disclosure to confidential to anonymous. (799)** **Since evaluation functions as an intervention, evaluators must strive to make their effect a positive one. (810)**

WHO EVALUATES: *who has a voice in the design of an evaluation, and who should perform evaluations*	
Sub-categories	Text of statements (edited form)
stakeholder input (29) standards for professionals (42) who collects data (1) who evaluates (60)	**The perspectives of diverse stakeholders shall be considered and included in evaluation design, implementation, analysis and reporting. (83)** **Require that evaluators be certified in order to perform evaluations of publicly funded programs. (96)** **Open task order qualifications on a frequent and regular basis to include newer firms in opportunities to compete for federal evaluation contracts. (124)** **When an evaluation team approach is adopted, the evaluation team shall include both content experts and professional evaluators. (276)** **Evaluation reports must explicitly acknowledge the ways in which major stakeholders have defined "success". (749)** **The government shall periodically assess the adequacy of the federal evaluation workforce. (756)** **Make the position of evaluator a career civil service one, instead of politically appointed. (778)** **Establish guidelines for when external evaluation is necessary. (790)** **To receive federal funding, an evaluator's primary place of business must be in the United States. (819)** **Establish a "better business"-type consumer protection rating system for all evaluation companies, groups, and individuals working with publicly funded evaluations. (833)**

Appendix E-Final statement set with ID numbers and ratings

New ID#	Original ID#	Statement	Avg. Merit	Avg. Feasi-bility	Com-bined
1	499	Shared understandings of program goals and mechanisms shall be re-assessed periodically.	4.35	3.87	8.22
2	600	In an evaluation, the philosophical biases of the evaluator must be clearly identified.	3.84	3.22	7.06
3	96	Require that evaluators be certified in order to perform evaluations of publicly funded programs.	2.68	3.04	5.72
4	608	Review Government Performance and Results Act (GPRA) indicators at least every other year and ask the leadership of funded programs for their input.	3.52	3.48	7.00
5	314	Apply complex, systems-based evaluation methods in complex, adaptive systems.	3.48	3.00	6.48
6	765	A "chief program evaluation officer" (CPEO) shall be appointed for each federal agency, and shall serve on a council with reporting requirements to Congress and the President.	3.37	3.13	6.50
7	749	Evaluation reports must explicitly acknowledge the ways in which major stakeholders have defined "success".	3.87	3.74	7.61
8	827	Evaluation findings and costs associated with publicly funded projects shall be fully disclosed in a manner accessible to the public.	4.26	3.74	8.00
9	422	Scoring rubrics including evaluation plans shall be published in advance as a part of RFP application review.	3.71	3.57	7.28
10	710	Establish guidelines for appropriate agency review of evaluation reports in order to protect the independence and objectivity of evaluators.	3.94	3.87	7.81
11	328	Data collection instruments shall be translated into languages other than English to ensure that all voices are heard.	3.23	3.30	6.53
12	833	Establish a "better business"-type consumer protection rating system for all evaluation companies, groups, and individuals working with publicly funded evaluations.	2.77	2.78	5.55
13	259	Encourage and value systematic internal evaluation by providing funding and bonus points to applicants with a history of focus on internal evaluation.	3.63	3.50	7.13
14	363	Evaluators shall serve as key members of the planning body for each project and program.	3.70	3.43	7.13
15	12	Value evaluation findings of "no discernable effect" as constructive feedback for program redirection.	4.13	3.65	7.78

16	814	Evaluations shall use many different ways to examine a question, including quantitative and qualitative, broad and narrow, shallow and deep.	4.19	3.83	8.02
17	32	Detect and decrease symbolic or pro-forma evaluations conducted solely to adhere in a minimum way to requirements for evaluation.	3.71	3.00	6.71
18	554	The federal government shall establish a clear, universal working definition of the term "program evaluation".	3.32	2.65	5.97
19	148	An evaluability assessment shall be conducted prior to launching into full-blown evaluation.	3.65	3.22	6.87
20	158	Standardize the evaluation language so that the evaluation work under one federal agency can be compared with the work under another agency.	3.55	2.57	6.12
21	774	Evaluations shall be required to use a random experimental design.	1.19	1.36	2.55
22	217	Value evaluation partnerships between academia and low-resource communities.	3.65	3.14	6.79
23	917	All federally funded programs must have a system in place for feedback and improvement.	4.26	3.65	7.91
24	123	Federally funded evaluations shall automatically provide access to data sets maintained by public entities, provided there is appropriate IRB oversight.	3.68	3.22	6.90
25	215	There shall be no comprehensive set of evaluation policies.	2.55	3.73	6.28
26	504	A contract for federal program evaluation shall be overseen by a branch or division that is over the entity being evaluated, not by the program that is the subject of the evaluation.	3.42	3.39	6.81
27	899	Federal reporting mechanisms shall be structured in such a way as to accommodate a variety of data collection and analysis methods.	4.42	3.82	8.24
28	498	Measure all outcomes, not just expected ones.	3.55	3.00	6.55
29	904	Organizations shall create logic models to guide their evaluations.	4.00	4.00	8.00
30	839	An independent agency shall be established to prioritize and systematically address attainment of pre-established goals and benchmarks of other federal agencies.	2.58	2.48	5.06
31	274	For all federally funded projects and programs, an evaluation plan must be established, documented and finalized within 60 days of funding.	3.10	2.30	5.40
32	896	There shall be a core set of evaluation policies which apply to all federal evaluations.	3.13	2.65	5.78
33	484	Establish guidelines for standards of evidence.	3.61	3.22	6.83

34	327	All federally funded projects shall include both process and outcome evaluation components.	3.48	3.00	6.48
35	776	Project managers must disclose political activities, lobbying, and funding from evaluation companies/individuals.	3.97	3.13	7.10
36	348	Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy.	3.77	3.52	7.29
37	611	Different ways of knowing from different cultures are valid as evidence	3.71	3.13	6.84
38	55	Rules and principles at the Federal level shall be devised in such a way as to set appropriate standards for evaluations at state and local levels.	3.23	2.91	6.14
39	86	Evaluations shall employ methods and procedures that fit the focus of the evaluation.	4.58	4.22	8.80
40	53	Build an evaluation culture.	4.06	3.35	7.41
41	591	Every federal program shall create a plan for how evaluations of grant initiatives will be used to inform decision making.	3.71	3.39	7.10
42	43	Evaluations shall employ measures that are reliable and valid for the respondents.	4.30	3.91	8.21
43	276	When an evaluation team approach is adopted, the evaluation team shall include both content experts and professional evaluators.	3.97	3.83	7.80
44	905	A toolbox or clearinghouse of tested evaluation methods and techniques shall be made available to programs required to complete evaluations.	3.71	3.65	7.36
45	810	Since evaluation functions as an intervention, evaluators must strive to make their effect a positive one.	3.19	2.83	6.02
46	10	Ideological philosophies shall be disclosed and articulated as underlying assumptions in program theory or logic.	3.45	2.74	6.19
47	898	Evaluation designs shall include principles of participatory action research, which involves all relevant parties in actively examining together some current action in order to change and improve it.	3.13	2.65	5.78
48	543	Evaluations are required to obtain an 80% response rate.	1.77	1.74	3.51
49	551	Establish guidelines for when and under what circumstances evaluations should be undertaken.	3.65	3.57	7.22
50	737	Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight.	3.73	3.52	7.25

51	114	All evaluations shall be informed by the most current knowledge about evaluation theory and practice.	3.84	2.96	6.80
52	132	A separate evaluation contract shall be included in major grants to ensure that funds available for evaluation are not reallocated to program efforts.	3.48	3.39	6.87
53	756	The government shall periodically assess the adequacy of the federal evaluation workforce.	3.50	3.09	6.59
54	708	Involve professional evaluators in reviewing the evaluation plans submitted on federal grants.	3.81	3.78	7.59
55	895	Make outcomes the primary focus of all evaluation plans.	2.48	2.96	5.44
56	181	Require that the timing of evaluation synchronize with program design and planning rhythms.	3.81	3.13	6.94
57	306	All evaluation reports conducted on federally funded evaluations shall be compiled in a central location that is open to the public.	3.87	3.22	7.09
58	394	Disaggregate evaluation results to inform policy in different settings.	3.61	3.00	6.61
59	83	The perspectives of diverse stakeholders shall be considered and included in evaluation design, implementation, analysis and reporting.	4.19	3.43	7.62
60	124	Open task order qualifications on a frequent and regular basis to include newer firms in opportunities to compete for federal evaluation contracts.	3.90	3.65	7.55
61	323	Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare.	4.39	3.96	8.35
62	77	Evaluation results around specific topic areas shall be published regularly by the federal government so as to identify the state of the art and any gaps in knowledge.	3.97	3.70	7.67
63	888	Replace the notion of generalizability with the recognition of local contextual realities as a basis of truth for national policy.	3.39	3.13	6.52
64	315	Only the English language shall be used in evaluation studies.	1.87	3.35	5.22
65	247	The standard for the evaluation of all social programs shall be whether the program advances the goals of social justice.	2.06	2.04	4.10
66	271	Evaluations shall employ methodologies that fit with program duration, intensity, lifecycle stage, context, available capacity, stakeholder needs.	4.45	3.57	8.02

67	643	Developmental evaluation approaches, which allow for the redirection of any evaluation as circumstances change, shall be considered for use in all new policy initiatives.	3.68	3.30	6.98
68	478	Where recommendations are enacted as a result of an evaluation, a second evaluation shall later be conducted to determine if they were successful.	3.45	2.91	6.36
69	604	Individual data gathered for one purpose may not be used for any other purpose without express permission of the individual.	3.13	3.39	6.52
70	534	Impact evaluations must utilize research designs that support strong causal inference.	3.65	3.09	6.74
71	26	Allow programs an appropriate period of time in operation before expecting them to achieve and report long-term, sustainable change.	4.55	4.09	8.64
72	599	The level of rigor of the evaluation shall be positively related to the magnitude and criticality of the decisions to be made on the basis of results.	4.13	3.96	8.09
73	575	Evaluation findings shall explicitly address threats to validity.	4.19	3.78	7.97
74	461	Every evaluation shall begin by identifying and articulating existing informal assessments and information feedback loops.	3.16	3.57	6.73
75	664	Evaluation findings shall include a discussion of the minority composition of the group studied and an indication of whether the general findings apply to minority groups.	3.06	3.30	6.36
76	790	Establish guidelines for when external evaluation is necessary.	3.90	3.74	7.64
77	602	Data analytic techniques employed in federal evaluations must be fully described and replicable.	4.29	3.65	7.94
78	485	Explicit criteria shall be developed for identifying "promising" innovative practices.	3.65	3.39	7.04
79	525	Evaluations shall be conducted with the goal of assuring that the population or program benefits from the evaluation.	3.71	3.22	6.93
80	210	Evaluate to learn not just what is happening but why.	4.19	3.48	7.67
81	359	A cost analysis (benefit, effectiveness, utility, and/or feasibility) shall be conducted for all federal programs.	3.00	2.48	5.48
82	727	There shall be a separate set of policies for research and development (R&D) operations.	3.27	3.52	6.79
83	778	Make the position of evaluator a career civil service one, instead of politically appointed.	3.55	3.35	6.90
84	137	All federally funded programs shall conduct regular fidelity assessments.	3.57	2.64	6.21

85	831	Establish guidelines for the level of funding required for evaluation conducted at varying levels of rigor, and using various methods.	3.84	3.09	6.93
86	746	Use evaluation continuously during the planning and management of federal government policies and activities, not just at budget time.	4.23	3.52	7.75
87	38	Evaluators reporting evidence of coercion through the exercise of power, influence, or resources shall be protected under Federal Whistle Blowers Laws.	4.27	3.57	7.84
88	740	Periodically undertake an evaluation of the federal program itself, not just its grantee programs.	4.35	3.52	7.87
89	799	Every evaluation study involving human subjects shall specify in advance the level of privacy for each phase of the study, based on a continuum from full disclosure to confidential to anonymous.	4.13	3.91	8.04
90	563	Where appropriate, agencies and programs shall use identical measures across programs and agencies.	3.42	2.61	6.03
91	808	Every evaluation plan shall indicate how the results are to be used and communicated.	4.23	3.74	7.97
92	436	Evaluations shall report not only majority views and findings but also minority views and findings.	3.87	3.39	7.26
93	74	Require training for federal, state and program managers--including what evaluation is, what constitutes effective evaluation work, and how to manage external evaluation.	4.06	3.48	7.54
94	8	Establish in advance measurable criteria by which the success of programs will be evaluated.	3.74	3.26	7.00
95	786	Require federal agencies to periodically publish draft versions of strategic plans and performance goals and indicators.	3.65	3.61	7.26
96	78	Support and nurture up-and-coming evaluators, through fellowships or sabbaticals at different agencies and opportunities for co-authorship.	3.77	3.61	7.38
97	832	Provide technical assistance to foster evaluation capacity building within organizations and agencies.	4.26	3.83	8.09
98	3	Grants shall include evaluation funding extending three or more years beyond the life of other grant funds, to assure follow-up.	3.71	2.74	6.45
99	819	To receive federal funding, an evaluator's primary place of business must be in the United States.	2.68	3.43	6.11
100	44	Require that evaluations be undertaken with the cultural "lens" appropriate to support decision making.	3.35	2.91	6.26

Appendix F: Statement listing by cluster with bridging values

Cluster 1: Relevance of reporting

58	Disaggregate evaluation results to inform policy in different settings.	.35
73	Evaluation findings shall explicitly address threats to validity.	.37
79	Evaluations shall be conducted with the goal of assuring that the population or program benefits from the evaluation.	.44
80	Evaluate to learn not just what is happening but why.	.44
1	Shared understandings of program goals and mechanisms shall be re-assessed periodically.	.62

Count:	**5**	Std. Dev.:	**0.09**	Minimum:	**0.35**	Average:	**.44**
		Variance:	**0.01**	Maximum:	**0.62**	Median:	**0.44**

Cluster 2: Respect for multiple perspectives

46	Ideological philosophies shall be disclosed and articulated as underlying assumptions in program theory or logic.	.53
51	All evaluations shall be informed by the most current knowledge about evaluation theory and practice.	.53
89	Every evaluation study involving human subjects shall specify in advance the level of privacy for each phase of the study, based on a continuum from full disclosure to confidential to anonymous.	.55
7	Evaluation reports must explicitly acknowledge the ways in which major stakeholders have defined "success".	.60
43	When an evaluation team approach is adopted, the evaluation team shall include both content experts and professional evaluators.	.60

Count:	**5**	Std. Dev.:	**0.03**	Minimum:	**0.53**	Average:	**.56**
		Variance:	**0.00**	Maximum:	**0.60**	Median:	**0.55**

Cluster 3: Justice of evaluation process

92	Evaluations shall report not only majority views and findings but also minority views and findings.	.62
75	Evaluation findings shall include a discussion of the minority composition of the group studied and an indication of whether the general findings apply to minority groups.	.64
65	The standard for the evaluation of all social programs shall be whether the program advances the goals of social justice.	.71
45	Since evaluation functions as an intervention, evaluators must strive to make their effect a positive one.	.77
2	In an evaluation, the philosophical biases of the evaluator must be clearly identified.	.88

Count:	**5**	Std. Dev.:	**0.10**	Minimum:	**0.62**	Average:	**.72**
		Variance:	**0.01**	Maximum:	**0.88**	Median:	**0.71**

Cluster 4: Respect for multiple methods

39	Evaluations shall employ methods and procedures that fit the focus of the evaluation.	.05
5	Apply complex, systems-based evaluation methods in complex, adaptive systems.	.08
16	Evaluations shall use many different ways to examine a question, including quantitative and qualitative, broad and narrow, shallow and deep.	.09

66 Evaluations shall employ methodologies that fit with program duration, intensity, lifecycle stage, context, available capacity, stakeholder needs. .09

42 Evaluations shall employ measures that are reliable and valid for the respondents. .09

28 Measure all outcomes, not just expected ones. .15

Count: **6** Std. Dev.: **0.03** Minimum: **0.05** Average: **.09**
Variance: **0.00** Maximum: **0.15** Median: **0.09**

Cluster 5: Strict standards for rigor

55 Make outcomes the primary focus of all evaluation plans. .19

19 An evaluability assessment shall be conducted prior to launching into full-blown evaluation. .21

21 Evaluations shall be required to use a random experimental design. .25

70 Impact evaluations must utilize research designs that support strong causal inference. .26

48 Evaluations are required to obtain an 80% response rate. .27

Count: **5** Std. Dev.: **0.03** Minimum: **0.19** Average: **.24**
Variance: **0.00** Maximum: **0.27** Median: **0.25**

Cluster 6: Requirements for evaluation

34 All federally funded projects shall include both process and outcome evaluation components. .27

94 Establish in advance measurable criteria by which the success of programs will be evaluated. .35

67 Developmental evaluation approaches, which allow for the redirection of any evaluation as circumstances change, shall be considered for use in all new policy initiatives. .36

29 Organizations shall create logic models to guide their evaluations. .46

Count: **4** Std. Dev.: **0.07** Minimum: **0.27** Average: **.36**
Variance: **0.00** Maximum: **0.46** Median: **0.35**

Cluster 7: Tailoring approach

37 Different ways of knowing from different cultures are valid as evidence for evaluation. .22

72 The level of rigor of the evaluation shall be positively related to the magnitude and criticality of the decisions to be made on the basis of results. .22

47 Evaluation designs shall include principles of participatory action research, which involves all relevant parties in actively examining together some current action in order to change and improve it. .29

74 Every evaluation shall begin by identifying and articulating existing informal assessments and information feedback loops. .31

100 Require that evaluations be undertaken with the cultural "lens" appropriate to support decision making. .32

59 The perspectives of diverse stakeholders shall be considered and included in evaluation design, implementation, analysis and reporting. .35

56 Require that the timing of evaluation synchronize with program design and planning rhythms. .36

11 Data collection instruments shall be translated into languages other than English to ensure that all voices are heard. .38

Count: **8** Std. Dev.: **0.06** Minimum: **0.22** Average: **.31**
Variance: **0.00** Maximum: **0.38** Median: **0.32**

Cluster 8: Certifying evaluator quality

62 Evaluation results around specific topic areas shall be published regularly by the federal government so as to identify the state of the art and any gaps in knowledge .44

12 Establish a "better business"-type consumer protection rating system for all evaluation companies, groups, and individuals working with publicly funded evaluations. .73

3 Require that evaluators be certified in order to perform evaluations of publicly funded programs. .83

Count: **3** Std. Dev.: **0.16** Minimum: **0.44** Average: **.67**
Variance: **0.03** Maximum: **0.83** Median: **0.73**

Cluster 9: Capacity building

40 Build an evaluation culture. .82

97 Provide technical assistance to foster evaluation capacity building within organizations and agencies. .87

96 Support and nurture up-and-coming evaluators, through fellowships or sabbaticals at different agencies and opportunities for co-authorship. .95

Count: **3** Std. Dev.: **0.05** Minimum: **0.82** Average: **.88**
Variance: **0.00** Maximum: **0.95** Median: **0.87**

Cluster 10: Integration of planning &

14 Evaluators shall serve as key members of the planning body for each project and program. .35

91 Every evaluation plan shall indicate how the results are to be used and communicated. .40

69 Individual data gathered for one purpose may not be used for any other purpose without express permission of the individual. .59

Count: **3** Std. Dev.: **0.11** Minimum: **0.35** Average: **.44**
Variance: **0.01** Maximum: **0.59** Median: **0.40**

Cluster 11: Guiding principles

15 Value evaluation findings of "no discernable effect" as constructive feedback for program redirection. .58

50 Value evaluation that does not facilitate programmatic decisions but is useful for learning and oversight. .70

22 Value evaluation partnerships between academia and low-resource communities. .87

36 Evaluators must adhere to the evaluation standards as developed by the Joint Committee on Evaluation Standards at the Evaluation Center (JSEC), Western Michigan University, on utility, feasibility, propriety, and accuracy. .96

61 Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare. 1.00

Count: **5** Std. Dev.: **0.16** Minimum: **0.58** Average: **.82**
Variance: **0.03** Maximum: **1.00** Median: **0.87**

Cluster 12: Openness and democracy

54 Involve professional evaluators in reviewing the evaluation plans submitted on federal grants. .07

95 Require federal agencies to periodically publish draft versions of strategic plans and performance goals and indicators. .07

60 Open task order qualifications on a frequent and regular basis to include newer firms in opportunities to compete for federal evaluation contracts. .11

8 Evaluation findings and costs associated with publicly funded projects shall be fully disclosed in a manner accessible to the public. .13

98 Grants shall include evaluation funding extending three or more years beyond the life of other grant funds, to assure follow-up. .13

Count: **5** Std. Dev.: **0.03** Minimum: **0.07** Average: **.10**
Variance: **0.00** Maximum: **0.13** Median: **0.11**

Cluster 13: Roles and relations

35 Project managers must disclose political activities, lobbying, and funding from evaluation companies/individuals. .18

83 Make the position of evaluator a career civil service one, instead of politically appointed. .21

93 Require training for federal, state and program managers--including what evaluation is, what constitutes effective evaluation work, and how to manage external evaluation. .25

13 Encourage and value systematic internal evaluation by providing funding and bonus points to applicants with a history of focus on internal evaluation. .33

87 Evaluators reporting evidence of coercion through the exercise of power, influence, or resources shall be protected under Federal Whistle Blowers Laws. .38

99 To receive federal funding, an evaluator's primary place of business must be in the United States. .39

Count: **6** Std. Dev.: **0.08** Minimum: **0.18** Average: **.29**
Variance: **0.01** Maximum: **0.39** Median: **0.29**

Cluster 14: Assuring use of results

86 Use evaluation continuously during the planning and management of federal government policies and activities, not just at budget time. .23

24 Federally funded evaluations shall automatically provide access to data sets maintained by public entities, provided there is appropriate IRB oversight. .27

41 Every federal program shall create a plan for how evaluations of grant initiatives will be used to inform decision making. .31

Count: **3** Std. Dev.: **0.03** Minimum: **0.23** Average: **.27**
Variance: **0.00** Maximum: **0.31** Median: **0.27**

Cluster 15: Institutionalizing evaluation

52 A separate evaluation contract shall be included in major grants to ensure that funds available for evaluation are not reallocated to program efforts. .00

53 The government shall periodically assess the adequacy of the federal evaluation workforce. .02

30 An independent agency shall be established to prioritize and systematically address attainment of pre-established goals and benchmarks of other federal agencies. .02

9 Scoring rubrics including evaluation plans shall be published in advance as a part of RFP application review. .03

88 Periodically undertake an evaluation of the federal program itself, not just its grantee programs. .03

4 Review Government Performance and Results Act (GPRA) indicators at least every other year and ask the leadership of funded programs for their input. .05

57 All evaluation reports conducted on federally funded evaluations shall be compiled in a central location that is open to the public. .06

26 A contract for federal program evaluation shall be overseen by a branch or division that is over the entity being evaluated, not by the program that is the subject of the evaluation. .07

23 All federally funded programs must have a system in place for feedback and improvement. .09

10 Establish guidelines for appropriate agency review of evaluation reports in order to protect the independence and objectivity of evaluators. .13

6 A "chief program evaluation officer" (CPEO) shall be appointed for each federal agency, and shall serve on a council with reporting requirements to Congress and the President. .14

Count: **11** Std. Dev.: **0.04** Minimum: **0.00** Average: **.06**
Variance: **0.00** Maximum: **0.14** Median: **0.05**

Cluster 16: Universal standards

20 Standardize the evaluation language so that the evaluation work under one federal agency can be compared with the work under another agency. .01

32 There shall be a core set of evaluation policies which apply to all federal evaluations .04

18 The federal government shall establish a clear, universal working definition of the term "program evaluation". .05

38 Rules and principles at the Federal level shall be devised in such a way as to set appropriate standards for evaluations at state and local levels. .06

31 For all federally funded projects and programs, an evaluation plan must be established, documented and finalized within 60 days of funding. .11

Count: **5** Std. Dev.: **0.03** Minimum: **0.01** Average: **.05**
Variance: **0.00** Maximum: **0.11** Median: **0.05**

Cluster 17: Aligning lifecycles

71 Allow programs an appropriate period of time in operation before expecting them to achieve and report long-term, sustainable change. .18

17 Detect and decrease symbolic or pro-forma evaluations conducted solely to adhere in a minimum way to requirements for evaluation. .19

Count: **2** Std. Dev.: **0.00** Minimum: **0.18** Average: **.18**
Variance: **0.00** Maximum: **0.19** Median: **0.18**

Cluster 18: Quality of evaluation practice

78 Explicit criteria shall be developed for identifying "promising" innovative practices. .22

68 Where recommendations are enacted as a result of an evaluation, a second evaluation shall later be conducted to determine if they were successful. .23

64 Only the English language shall be used in evaluation studies. .26

63 Replace the notion of generalizability with the recognition of local contextual realities as a basis of truth for national policy. .28

77 Data analytic techniques employed in federal evaluations must be fully described and replicable. .30

Count: **5** Std. Dev.: **0.03** Minimum: **0.22** Average: **.26**
Variance: **0.00** Maximum: **0.30** Median: **0.26**

Cluster 19: Guidelines

82 There shall be a separate set of policies for research and development (R&D) operations. .06

25 There shall be no comprehensive set of evaluation policies. .14

84 All federally funded programs shall conduct regular fidelity assessments. .15

81 A cost analysis (benefit, effectiveness, utility, and/or feasibility) shall be conducted for all federal programs. .16

85 Establish guidelines for the level of funding required for evaluation conducted at varying levels of rigor, and using various methods. .23

49 Establish guidelines for when and under what circumstances evaluations should be undertaken. .25

76 Establish guidelines for when external evaluation is necessary. .26

Count: **7** Std. Dev.: **0.07** Minimum: **0.06** Average: **.18**
Variance: **0.00** Maximum: **0.26** Median: **0.16**

Cluster 20: Communicating and

90 Where appropriate, agencies and programs shall use identical measures across programs and agencies. .23

27 Federal reporting mechanisms shall be structured in such a way as to accommodate a variety of data collection and analysis methods. .35

44 A toolbox or clearinghouse of tested evaluation methods and techniques shall be made available to programs required to complete evaluations. .41

33 Establish guidelines for standards of evidence. .51

Count: **4** Std. Dev.: **0.10** Minimum: **0.23** Average: **.37**
Variance: **0.01** Maximum: **0.51** Median: **0.38**

Appendix G-Top 20 “go-zone” statements

cluster	#	statement,	combined rating (out of 10)
1	UTILITY OF RESULTS		
	1	Shared understandings of program goals and mechanisms shall be re-assessed periodically.	8.40
	73	Evaluation findings shall explicitly address threats to validity.	8.07
	80	Evaluate to learn not just what is happening but why.	7.82
2	RESPECT FOR MULTIPLE PERSPECTIVES		
	7	Evaluation reports must explicitly acknowledge the ways in which major stakeholders have defined "success".	8.20
	89	Every evaluation study involving human subjects shall specify in advance the level of privacy for each phase of the study, based on a continuum from full disclosure to confidential to anonymous.	8.06
4	RESPECT FOR MULTIPLE METHODS		
	39	Evaluations shall employ methods and procedures that fit the focus of the evaluation.	8.87
	66	Evaluations shall employ methodologies that fit with program duration, intensity, lifecycle stage, context, available capacity, stakeholder needs.	8.18
	16	Evaluations shall use many different ways to examine a question, including quantitative and qualitative, broad and narrow, shallow and deep.	8.15
	42	Evaluations shall employ measures that are reliable and valid for the respondents.	8.09
5	STRICT STANDARDS FOR RIGOR		
6	REQUIREMENTS FOR EVALUATION PLANS		
	29	Organizations shall create logic models to guide their evaluations.	7.77
8	CERTIFYING EVALUATOR QUALITY		
	62	Evaluation results around specific topic areas shall be published regularly by the federal government so as to identify the state of the art and any gaps in knowledge.	7.80
9	CAPACITY BUILDING		
	97	Provide technical assistance to foster evaluation capacity building within organizations and agencies.	8.33
10	ETHICAL USE OF RESULTS		
	91	Every evaluation plan shall indicate how the results are to be used and communicated.	8.18
11	GUIDING PRINCIPLES		

	61	Evaluators must adhere to American Evaluation Association (AEA) Guiding Principles for Evaluators on systematic inquiry, competence, integrity/honesty, respect for people, and responsibilities for general and public welfare.	8.32
	15	Value evaluation findings of "no discernable effect" as constructive feedback for program redirection.	7.78
12		OPENNESS AND DEMOCRACY	
	8	Evaluation findings and costs associated with publicly funded projects shall be fully disclosed in a manner accessible to the public.	8.18
13		ROLES AND RELATIONS	
	87	Evaluators reporting evidence of coercion through the exercise of power, influence, or resources shall be protected under Federal Whistle Blowers Laws.	8.27
14		EVALUATOR EFFECTIVENESS	
	86	Use evaluation continuously during the planning and management of federal government policies and activities, not just at budget time.	7.83
15		INSTITUTIONALIZING EVALUATION	
	23	All federally funded programs must have a system in place for feedback and improvement.	8.06
	88	Periodically undertake an evaluation of the federal program itself, not just its grantee programs.	8.00
	10	Establish guidelines for appropriate agency review of evaluation reports in order to protect the independence and objectivity of evaluators.	7.89
16		UNIVERSAL STANDARDS	
17		ALLIGNING TO PROGRAM LIFECYCLE	
	71	Allow programs an appropriate period of time in operation before expecting them to achieve and report long-term, sustainable change.	8.52
18		QUALITY OF EVALUATION PRACTICE	
	77	Data analytic techniques employed in federal evaluations must be fully described and replicable.	8.03
19		GUIDELINES	
	76	Establish guidelines for when external evaluation is necessary.	7.82
20		COMMUNICATING AND COORDINATING	
	27	Federal reporting mechanisms shall be structured in such a way as to accommodate a variety of data collection and analysis methods.	8.22

SOURCES:

Alkin, Marvin C. and Fred Ellett (1990) Evaluation models, in Walberg, H. J., & Haertel, G. D. (1990). *The international encyclopedia of educational evaluation*. Oxford, England; New York: Pergamon Press.

American Educational Research Association (http://www.aera.net).

American Evaluation Association (2007) *Internal scan report to the membership*, prepared by Goodman Research Group, Inc. submitted April 2008.AEA Newsletter July 2011, Vol. 11, Issue 7, "Policy Watch" by George Grob. (http://www.eval.org/)

American Evaluation Association (2008), letter from President Bill Trochim to Robert Shea, Associate Director for Administration and Government Performance, United States Office of Management and Budget, 3/07/08.

American Evaluation Association: Guiding principles for evaluators. (2009). *American Journal of Evaluation, 30*(3), 273-274.

American Evaluation Association: An evaluation roadmap for a more effective government. (2009). Prepared by the *AEA Evaluation Policy Task Force*, 2/2009 and 10/2010. http://www.eval.org/aea09.eptf.eval.roadmapF.pdf (retrieved 5/15/10 and 9/1/11).

American Evaluation Association (2010) mission statement, http://www.eval.org/aboutus/organization/aboutus.asp (retrieved 5/10/10).

Balthasar, A. (2009). Institutional design and utilization of evaluation: A contribution to a theory of evaluation influence based on swiss experience. *Evaluation Review, 33*(3), 226-256.

Bemelmans-Videc, M., Rist, R. C., & Vedung, E.,. (1998). *Carrots, sticks & sermons : Policy instruments and their evaluation*. New Brunswick, N.J., U.S.A.: Transaction Publishers.

Blumer, H. (1954). What is wrong with social theory? *American Sociological Review, 19*(1), 3-10.

Caracelli, V. J. (1989). Structured conceptualization: A framework for interpreting evaluation results.*12*(1), 45-52.

Carman, J. G., Fredericks, K. A., & Introcaso, D. (2008). Government and accountability: Paving the way for nonprofits and evaluation. *New Directions for Evaluation,* (119), 5-12.

Campbell, D. T. (1966). Pattern matching as an essential in distal knowing. In Hammond, K. R., (1966). *The psychology of egon brunswik,*. New York: Holt, Rinehart and Winston.

Cañas, A. J., Novak, J. D., & González García, F. M. (2004). *Concept maps : Theory, methodology, technology : Proceedings of the first international conference on concept mapping, CMC 2004, pamplona, spain, sept 14-17, 2004*. [Pamplona: Dirección de Publicaciones de la Universidad Pública de Navarra.

Carver, J., & Carver, M. M. (1997). *Reinventing your board: A step-by-step guide to implementing policy governance. the jossey-bass nonprofit and public management series*. Jossey-Bass Inc., Publishers, 350 Sansome Street, San Francisco, CA 94104 ($32). Tel: 415-433-1740; Fax: 415-433-0499; Web site: http://www.josseybass.com; e-mail: webperson@jbp.com.

Chamarz, Kathy (2001) Grounded theory in Smith, J. A., Harré, R., & Langenhove, L. v. (1995). *Rethinking methods in psychology*. London; Thousand Oaks, Calif.: Sage Publications.

Chelimsky, E. (1985) Old patterns and new directions in program evaluation. In Chelimsky, E., & American Society for Public Administration. (1985). *Program evaluation : Patterns and directions*. Washington, D.C.: American Society for Public Administration.

Chelimsky, E. (1986) What have we learnt about the politics of program evaluation? Plenary address to the American Evaluation Association, Kansas, Missouri.

Chelimsky, E. (2007). Factors influencing the choice of methods in federal evaluation practice.(113), 13-33.

Chelimsky, E. (2009). Integrating evaluation units into the political environment of government: The role of evaluation policy.(123), 51-66. doi:10.1002/ev.305

Coalition for Evidence-Based Policy (http://coalition4evidence.org/wordpress/)

Compton, D. W., Glover-Kudon, R., Smith, I. E., & Eden Avery, M. (2002). Ongoing capacity building in the american cancer society (ACS). *New Directions for Evaluation, 2002*(93), 47-62. doi:10.1002/ev.41

Compton, D., Baizerman, M., & Hueftle Stockdill, S. (2002). The art, craft, and science of evaluation capacity building. *New Directions for Evaluation, 2002*(93), i.

Concept Systems Incorporated (2005) http://www.conceptsystems.com/

Constas, M. A. (1992). Qualitative analysis as a public event: The documentation of category development procedures. *American Educational Research Journal, 29*(2), 253-266.

Cook, C., Heath, F., & Thompson, R. L. (2000). A meta-analysis of response rates in web- or internet-based surveys. *Educational and Psychological Measurement, 60*(6), 821-836. doi:10.1177/00131640021970934

Cooksy, L. J., Mark, M. M., & Trochim, W. M. K. (2009). Evaluation policy and evaluation practice: Where do we go from here? *New Directions for Evaluation, 2009*(123), 103-109. doi:10.1002/ev.308

Cooper, H. M. (1988). Organizing knowledge syntheses: A taxonomy of literature reviews. *Knowledge in Society, 1*(1), 104-126.

Cordray, D., & Lipsey, M. (1986). Introduction, evaluation studies in 1986: Program evaluation and program research. *Evaluation studies review annual* (11), 17-44).

Cronbach, L. J. (1963) Course improvement through evaluation in *Teachers' College Record* (64), 672–83.

Cousins, J. B., & Whitmore, E. (2007). Framing participatory evaluation.(114), 87-105. doi:10.1002/ev

Dahler-Larsen, Peter (2006) Organizing knowledge: evidence and the construction of evaluative information systems. In Rist, R. C., & Stame, N. (2006). *From studies to streams : Managing evaluative systems : Comparative policy evaluation, volume XII* Transaction Publishers.

Datta L.-e. (2011). Politics and evaluation: More than methodology. *American Journal of Evaluation, 32*(2), 273-294.

Datta, L. (2009). Golden is the sand: Memory and hope in evaluation policy and evaluation practice.(123), 33-50. doi:10.1002/ev.304

Debelstein, Niels and C. Rebien (2002) Evaluation of Developmental Assistance: Its Start, Progress and Current Challenges. In *International atlas of evaluation*. New Brunswick, U.S.A.: Transaction Publishers.

Eisner, E. W. (1979). *The educational imagination : On the design and evaluation*. New York: Macmillan.

Everitt, B. In Social Science Research Council (Great Britain) (Ed.), *Cluster analysis* (2d ed. ed.) London : published on behalf of the Social Science Research Council by Heinemann Educational Books ; New York : Halsted Press, 1980.

Foresti, Marta with Archer, C., O'Neil, T., Longhurst, R. (2007) A comparative study of evaluation policies and practices in development agencies, Research Department, Division of Evaluation and Capitalization, Overseas Development Institute, Methodological Series, No. 1, Paris, France. Retrieved from the Overseas Development Institute (http://www.odi.org.uk/resources/details.asp?id=3320&title=evaluation-policy-practice-development-aid-agencies)

Furubo, J., Rist, R. C., & Sandahl, R. (2002). *International atlas of evaluation*. New Brunswick, U.S.A.: Transaction Publishers.

Garms, W. I., Guthrie, J. W., & Pierce, L. C.,. (1978). *School finance : The economics and politics of public education*. Englewood Cliffs, N.J.: Prentice-Hall.

Geva-May, I., & Pal, L. A. (1999). Good fences make good neighbours: Policy evaluation and policy analysis - exploring the differences. *Evaluation, 5*(3), 259.

Glaser, B. G., & Strauss, A. L.,. (1967). *The discovery of grounded theory; strategies for qualitative research*. Chicago: Aldine Pub. Co..

Government Performance and Results Act (1993) P.L. 103-62

Grasso, P. G. (1996). End of an era: Closing the U.S. general accounting office's program evaluation and methodology division. *Evaluation Practice, 17*(2), 115.

Greenberg, G. D., Miller, J. A., Mohr, L. B., & Vladeck, B. C. (1977). Developing public policy theory: Perspectives from empirical research. *The American Political Science Review, 71*(4), pp. 1532-1543.

Greene, J. C., Caracelli, V. J., & Graham, W. F. (1989). Toward a conceptual framework for mixed-method evaluation designs. *Educational Evaluation and Policy Analysis, 11*(3), pp. 255-274.

Grob, G. F. (2003). A truly useful bat is one found in the hands of a slugger. *American Journal of Evaluation, 24*(4), 499-505. doi:10.1177/109821400302400407

Grob, George, personal email to American Evaluation Association Evaluation Policy Task Force list serve, 2/22/10.

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding Evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314. doi:10.1177/109821400302400302

Henry, G. T., & Mark, M. M. (2003). Beyond use: Understanding Evaluation's influence on attitudes and actions. *American Journal of Evaluation, 24*(3), 293-314. doi:10.1177/109821400302400302

House, E. R. (1980). *Evaluating with validity*. Beverly Hills, Calif.: Sage Publications.

Jackson, K. M., & Trochim, W. M. K. (2002). Concept mapping as an alternative approach for the analysis of open-ended survey responses. *Organizational Research Methods, 5*(4), 307-336. doi:10.1177/109442802237114

Joint Committee on Standards for Educational Evaluation. (1981). *Standards for evaluations of educational programs, projects, and materials*. New York, N.Y.: McGraw-Hill Book Co.

Julnes, G., & Rog, D. J. (2007). Pragmatic support for policies on methodology. *New Directions for Evaluation, 2007*(113), 129-147. doi:10.1002/ev.219

Kane, M., & Trochim, W. M. K. (2007). *Concept mapping for planning and evaluation*. Thousand Oaks: Sage Publications.

King, J. A. (2002). Building the evaluation capacity of a school district. *New Directions for Evaluation, 2002*(93), 63-80. doi:10.1002/ev.42

Krippendorff, K. (1980). *Content analysis : An introduction to its methodology*. Beverly Hills: Sage Publications.

Kruskal, J. B., & Wish, M.,. (1978). *Multidimensional scaling*. Beverly Hills, Calif.: Sage Publications.

Lambur, M. T. (2008). Organizational structures that support internal program evaluation. *New Directions for Evaluation, 2008*(120), 41-54. doi:10.1002/ev.275

Leeuw F.L., & Furubo J.-E. (2008). Evaluation systems: What are they and why study them? *Evaluation Evaluation, 14*(2), 157-169+260.

Leeuw, F. L. (2009). Evaluation policy in the netherlands. *New Directions for Evaluation, 2009*(123), 87-102. doi:10.1002/ev.307

Lincoln, Y. S., & Guba, E. G. (1985). *Naturalistic inquiry*. Beverly Hills, Calif.: Sage Publications.

Liverani, A., & Lundgren, H. (2007). Evaluation systems in development aid agencies. *Evaluation, 13*(2), 241-256.

Liverani, A., & Lundgren, H. E. (2007). Evaluation systems in development aid agencies: An analysis of DAC peer reviews 1996-2004. *Evaluation: The International Journal of Theory, Research and Practice, 13*(2), 241-256. doi:10.1177/1356389007075226

Lowi, T. (1964). American business, public policy, case–studies, and political theory. *World Politics,* XVI (4), 677-715.

Mackay, K. R., & World Bank. Independent Evaluation Group. (2007). *How to build M & E systems to support better government.* Retrieved

Mark, M. M., Cooksy, L. J., & Trochim, W. M. K. (2009). Evaluation policy: An introduction and overview. *New Directions for Evaluation,* (123), 3-11. doi:10.1002/ev.302

McCallin A. (2003). Grappling with the literature in a grounded theory study. *Contemporary Nurse : A Journal for the Australian Nursing Profession, 15*(1-2), 1-2.

McDavid, J. C., & Hawthorn, L. R. L. (2005). *Program evaluation and performance measurement : An introduction to practice*. London: SAGE.

Merriam Webster Online (http://www.merriam-webster.com/dictionary/policy) retrieved 5/11/10.

Mitchell, D. E., & Iannaccone, L.,. (1980). *The impact of California's legislative policy on public school performance*. Berkeley: Institute of Governmental Studies, University of California.

Mitchell, D. E., Marshall, C., & Wirt, F. M. (1985). Building a taxonomy of state education policies. *Peabody Journal of Education, 62*(4), 7-47.

Nagel, I. H., Freeman, H. E., & Russell Sage Foundation. (1975). *Academic and entrepreneurial research : The consequences of diversity in federal evaluation studies*. New York: Russell Sage Foundation.

Nelson, E., Bemelmans-Vidce, M., Rist, R., & Vedung, E. (1998). Carrots, sticks, and sermons: Policy instruments and their evaluation. *Knowledge, Technology and Policy, 11*(3), 68-69.

Novak, J. D., & Gowin, D. B. (1984). *Learning how to learn*. Cambridge [Cambridgeshire]; New York: Cambridge University Press.

Organisation for Economic Co-operation and Development. Development Assistance Committee. (1992). *DAC principles for effective aid : Development assistance manual.* Paris; Washington, D.C.: Organisation for Economic Co-operation and Development ; OECD Publications and Information Centre, distributor].

Organisation for Economic Co-operation and Development. Development Assistance Committee. (1998). *Review of the DAC principles for evaluation of development assistance*. Paris: OECD.

Organisation for Economic Co-operation and Development. Development Assistance Committee. (2004) "Evaluation Systems in DAC Members Agencies", a study based on DAC Peer Reviews, presented at the Second Meeting of the DAC Network on Development Evaluation, Paris, 9 October 2004, p.4. (http://www.oecd.org)

Phi Delta Kappa. National Study Committee on Evaluation. (1971). *Educational evaluation & decision making.* Itasca, Ill.: F.E. Peacock Publishers.

Rist, Ray and Paliokas, K. (2002) The rise and fall (and rise again) of the evaluation function in the U.S. government. In *International atlas of evaluation*. New Brunswick, U.S.A.: Transaction Publishers.

Rosas, S. R., & Kane, M. (2011). Quality and rigor of the concept mapping methodology: A pooled study analysis. *Evaluation and Program Planning, 35*(2), 236-245. doi:10.1016/j.evalprogplan.2011.10.003

Rossi, P. H., Lipsey, M. W., & Freeman, H. E. (2004). *Evaluation : A systematic approach*. Thousand Oaks, CA: Sage.

Russon, C. (2004). Cross-cutting issues in international standards development. *New Directions for Evaluation, 2004*(104), 89-93. doi:10.1002/ev.139

Sanders, J. R. (1994). *The process of developing national standards that meet ANSI guidelines.*

Schaumburg-Müller, H. (2005). Use of aid evaluation from an organizational perspective. *Evaluation, 11*(2), 207-222. doi:10.1177/1356389005055541

Scheirer, M. A., & Newcomer, K. (2001). Opportunities for program evaluators to facilitate performance-based management. *Evaluation and Program Planning, 24*(1), 63-71.

Scriven, M. (1998). Minimalist theory: The least theory that practice requires. *American Journal of Evaluation, 19*(1), 57-70. doi:10.1177/109821409801900105

Scriven, M. (1966). *The methodology of evaluation*. Lafayette, Ind.: Purdue University.

Segerholm, C. (2003). Researching evaluation in national (state) politics and administration: A critical approach. *American Journal of Evaluation, 24*(3), 353-372. doi:10.1177/109821400302400305

Segerholm, C. (2003). Researching evaluation in national (state) politics and administration: A critical approach. *American Journal of Evaluation, 24*(3), 353-372. doi:10.1177/109821400302400305

Segsworth, R. V. (2005). Program evaluation in the government of canada: Plus ca change. *CANADIAN JOURNAL OF PROGRAM EVALUATION, 20*(3), 175-198.

Shadish, W. R. (1998). Evaluation theory is who we are. *American Journal of Evaluation, 19*(1), 1-19. doi:10.1177/109821409801900102

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). *Foundations of program evaluation : Theories of practice*. Newbury Park, Calif.: Sage Publications.

Small, M. (2009). 'How many cases do I need?': On science and the logic of case selection in field-based research. *Ethnography, 10*(1), 5-38. doi:10.1177/1466138108099586

Smith, K. B. (2002). Typologies, taxonomies, and the benefits of policy classification. *Policy Studies Journal, 30*(3), 379.

Stemler, S. (2001). An overview of content analysis. *Practical Assessment, Research & Evaluation, 7*(17)

Stern, E. (2009). Evaluation policy in the European Union and its institutions. *New Directions for Evaluation, 2009*(123), 67-85.

Stevahn, L., King, J. A., Ghere, G., & Minnema, J. (2005). Establishing essential competencies for program evaluators. *American Journal of Evaluation, 26*(1), 43-59. doi:10.1177/1098214004273180

Strauss, A. L., & Corbin, J. M.,. (1990). *Basics of qualitative research : Grounded theory procedures and techniques*. Newbury Park, Calif.: Sage Publications.

Stufflebeam, D. L., Madaus, G. F., & Kellaghan, T. (2000). *Evaluation models : Viewpoints on educational and human services evaluation*. Boston: Kluwer Academic Publishers.

Summa, Hilkka and Toulemond, Jacques (2002) Evaluation in the European Union: addressing complexity and ambiguity, pp. 207-424. In *International atlas of evaluation*. New Brunswick, U.S.A.: Transaction Publishers.

Tashakkori, A., & Teddlie, C. (1998). *Mixed methodology : Combining qualitative and quantitative approaches*. Thousand Oaks, Calif.: Sage.

Taylor-Powell, E., & Boyd, H. H. (2008). Evaluation capacity building in complex organizations. *New Directions for Evaluation, 2008*(120), 55-69. doi:10.1002/ev.276

Treasury Board of Canada (2009) *Policy on evaluation* (http://www.tbs-sct.gc.ca/pol/doc-eng.aspx?section=text&id=15024) retrieved 5/15/10

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review, 9*(5), 575-604. doi:10.1177/0193841X8500900503

Trochim, WM & Linton, R. (1986). Conceptualization for planning and evaluation. *Evaluation and Program Planning, 9*(4), 289-308.

Trochim, W. (1989). An introduction to concept mapping for planning and evaluation. *Evaluation and Program Planning, 12*(1), 1-16.

Trochim, W. M. K. (1985). Pattern matching, validity, and conceptualization in program evaluation. *Evaluation Review Evaluation Review, 9*(5), 575-604.

Trochim, W.M.K. (1993) *The reliability of concept mapping*. Paper presented at the Annual Conference of the American Evaluation Association, Dallas, TX.

Trochim, W. M. K. (1998). An evaluation of michael Striven's "Minimalist theory: The least theory that practice requires". *American Journal of Evaluation, 19*(2), 243-249. doi:10.1177/109821409801900211

Trochim, W. M. K. (2009). Evaluation policy and evaluation practice.(123), 13-32. doi:10.1002/ev.303

Trochim, W. M. K. (2009). Evaluation policy and evaluation practice. *New Directions for Evaluation, 2009*(123), 13-32. doi:10.1002/ev.303

Trochim, William M. (2010) *Research Methods Knowledge Base,* http://www.socialresearchmethods.net/kb/

U.S. Congressional Research Service (2006) *Congress and program evaluation; an overview of randomized controlled trials (RCTs) and related issues*, a CRS report for Congress, order number RL 33301 by Clinton Brass, Blaz Nunez-Neto and Erin Williams. (http://opencrs.com/document/RL33301/2006-03-07/).

U.S. Congressional Research Service (2011) *Obama Administration Agenda for Government Performance: Evolution and Related Issues for Congress,* a CRS Memorandum prepared for distribution to Congress, by Clinton Brass.

(http://www.scribd.com/doc/48106516/CRS-Memo-on-Obama-Performance-Agenda-1-19-11)

U.S. Department of State (2011) *Department of State Evaluation Policy* (http://www.state.gov/s/d/rm/rls/fs/2011/163299.htm) retrieved 6/1/2011

U.S. Government Accountability Office (1997, June) the government performance and results act: 1997 governmentwide implementation will be uneven. (Publication no. GAO/GGD-97-109). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html.

U.S. Government Accountability Office (1998, April) Agencies Challenged by New Demand for Information on Program Results (Publication No. GAO/GGD-98-53). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html

U.S. Government Accountability Office (2000a, July) managing for results; continuing challenges to effective GPRA implementation. (publication no. GAO/T-GGD-00-178). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html.

U.S. Government Accountability Office (2000b, March) managing for results; barriers to interagency coordination. (publication no. GAO/GGD-00-106). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html.

U.S. Government Accountability Office (2004, January) performance budgeting observations on the use of OMB's program, assessment rating tool for the fiscal year 2004 budget. (publication no. GAO-04-174). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html.

U.S. Government Accountability Office (2005, October) PART focuses attention on program performance, but more can be done to engage congress. (publication no. GAO-O6-28). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html

U.S. Government Accountability Office (2009, November) A variety of rigorous methods can help identify effective interventions. (publication no. GAO-10-30). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.html.

U.S. Government Accountability Office (2010, May) streamlining government; opportunities exist to strengthen OMB's approach to improving efficiency. (publication no. GAO-10-394). retrieved from GAO reports main page via GPO access database: http://www.gpoaccess.gov/gaoreports/index.htm

U.S. Office of Management and Budget (2004) *What Constitutes Strong Evidence of a Program's Effectiveness?*

(http://www.whitehouse.gov/sites/default/files/omb/part/2004_program_eval.pdf)

U.S. Office of Management and Budget, Oct. 7, 2009 Memorandum of OMB Director Peter Orzag to Heads of Executive Departments and Agencies, entitled "Increased Emphasis on Program Evaluations". M-10-01/ (http://www.whitehouse.gov/sites/default/files/omb/assets/memoranda_2010/m10-01.pdf)

U.S. Presidential Budget (2010), Analytic Perspectives, Section 8 "Evaluation Initiative", United States Presidential Budget Proposal for 2011, *Analytical Perspectives: Budget of the U.S. Government*, (http://www.whitehouse.gov/omb/budget/fy2011/assets/spec.pdf) pages 71-98.

U.S. Presidential Budget (2011) *Analytic Perspectives: Budget of the U.S. Government, United States Presidential Budget Proposal for 2012,* (http://www.whitehouse.gov/omb/budget/Analytical_Perspectives/)

Rupsiene, L. & Pranskuniene, R. (2010) The variety of grounded theory: Different versions of the same method or different methods? *Social Sciences (1392-0758), 70*(4), 7-19.

Walberg, H. J., & Haertel, G. D. (1990). *The international encyclopedia of educational evaluation*. Oxford, England; New York: Pergamon Press.

Ward, J. H. (1963). Hierarchical grouping to optimize an objective function. *Journal of the American Statistical Association, 58*(301), 236-244.

Weber, R. P. (1990). *Basic content analysis*. Newbury Park, Calif.: Sage Publications.

Weiss, R. S.,. (1994). *Learning from strangers : The art and method of qualitative interview studies*. New York; Toronto; New York: Free Press ; Maxwell Macmillan Canada ; Maxwell Macmillan International.

Weiss, C. H. (1997). Theory-based evaluation: Past, present, and future. *New Directions for Evaluation, 1997*(76), 41-55. doi:10.1002/ev.1086

Weiss, C. H., Murphy-Graham, E., & Birkeland, S. (2005). An alternate route to policy influence. *American Journal of Evaluation, 26*(1), 12-30. doi:10.1177/1098214004273337

Weller, S. C., & Romney, A. K. (1988). *Systematic data collection*. Newbury Park, Calif.: Sage Publications.

Wholey, J. S. (1970). *Federal evaluation policy; analyzing the effects of public programs,*. Washington: Urban Institute.

Wholey, J. S. (2001). Managing for results: Roles for evaluators in a new management era. *American Journal of Evaluation, 22*(3), 343-347. doi:10.1177/109821400102200309

Wisler, C. E., & American Evaluation Association. (1996). *Evaluation and auditing : Prospects for convergence*. San Francisco, Calif.: Jossey-Bass Pub.

Worthen, B. R. (1997). In Sanders, James R.Fitzpatrick, Jody LWorthen,Blaine R.Educational evaluation. (Ed.), *Program evaluation : Alternative approaches and practical guidelines* (2nd ed. ed.). New York: Longman.

Yarbrough, D. B., Shulha, L. M., & Caruthers, F. (2004). Background and history of the joint committee's program evaluation standards. *New Directions for Evaluation, 2004*(104), 15-30. doi:10.1002/ev.133