AMERICAN EVALUATION ASSOCIATION

March 7, 2008

Robert Shea
Associate Director for Administration and Government Performance
Office of Management and Budget
Eisenhower Executive Office Building
Washington, DC

Dear Mr. Shea,

I am pleased to submit, in response to your request on February 26, 2008, the comments of the Evaluation Policy Task Force of the American Evaluation Association (AEA) on the evaluation guidelines represented in the document entitled "What Constitutes Strong Evidence of a Program's Effectiveness?"

We support efforts to implement systematic processes for evaluating federal programs and we see the OMB Program Assessment Rating Tool (PART) as an important step in that direction. With appropriate revisions to the existing evaluation guidelines, the PART could be even more useful, reliable, and sustainable. PART is well constructed to encourage a comprehensive evaluation approach for the diversity of federal programs.

Our primary overall recommendations are:

- **Develop new guidance around evaluation in PART Questions 2.6 and 4.6:** Clarifying and expanding the evaluation guidance for PART Questions 2.6 and 4.5 would improve PART's legacy, sustainability, and validity. To provide practical guidance for OMB examiners and agency officials, such a document would include:

  a) How evaluation information from other PART questions on program purpose and design, program strategic planning and measurement, and program management is needed and used as a foundation for designing high quality effectiveness evaluation.

  b) How to identify and select evaluation methods for assessing program effectiveness that are appropriate to the needs and development level of a program.

- **Present Randomized Control Trials (RCTs) in context:** More balanced and considered presentation of the role of RCTs in assessing the effectiveness of federal programs would significantly improve the document.

  a) There are important alternatives to RCTs for assessing effectiveness that should be included and may have distinct advantages under specific circumstances.

16 Sconticut Neck Rd #290, Fairhaven MA 02719 USA • Phone 1-508-748-3326 Fax 1-508-748-3158 • info@eval.org www.eval.org

b) Situating RCTs, the focus of the current document, within a balanced presentation of the spectrum of appropriate, rigorous, evaluation methods will improve the likelihood of selecting appropriate measures and methods to assess and improve program performance.

c) The utility of RCTs can be significantly enhanced when mixed with additional methods that enable identification of why and how observed effects occur.

- **Draw on a range of expertise:** Calling upon experts who could provide relevant assistance during the development of the evaluation guidance would enhance the knowledge-base of the drafters and improve the rigor and feasibility of the guidance. We would be interested in either working with you directly on revisions of the guidance or in working with you and your staff to identify appropriate evaluation expertise.

Given the short timeframe, we have not been able to include detailed citations and references to the literature. Should such additional be beneficial, please do not hesitate to contact us.

We hope that our comments are helpful to you and your staff at OMB and we look forward to working with you to improve the PART system so that it is both effective and sustainable as a management mechanism in the federal government.

Sincerely,

William M.K. Trochim
President, American Evaluation Association
Chair, AEA Evaluation Policy Task Force

Attachments:

Comments on “What Constitutes Strong Evidence of a Program’s Effectiveness?”

Comments on
What Constitutes Strong Evidence of a Program's Effectiveness?

Evaluation Policy Task Force
American Evaluation Association
March 7, 2008

The Evaluation Policy Task Force supports the general idea behind the PART process, and particularly its recognition of the important role that program evaluation can play in improving program performance and accountability. We recognize the hard work that has gone into developing the guidance to date and commend OMB for building on their investment by continually improving the PART program in critically important ways.

The PART process presents an opportunity to promote high-quality program evaluation methodologies designed to assess and improve federal government programs. The paper at issue is useful in that it brings together considerable information on randomized experimental designs or Randomized Clinical Trials (RCTs). However, it has serious limitations that threaten the viability and sustainability of the PART process. We provide these comments in the hope that OMB will consider significant revision and enhancement of this statement, and are confident that addressing these concerns would strengthen the PART process and lead to improved federal program performance.

This paper misses a key opportunity to help federal programs align their most pressing evaluation questions with the most rigorous, practicable evaluation methodology. While the paper notes (p. 1) that "RCTs are not suitable for every program and generally can be employed only under very special circumstances," and thus protects itself from criticisms of advocating only one methodology, it subsequently neglects consideration of how agencies would determine where, when, and why other methods might be used. Instead, the paper focuses almost exclusively on RCTs.

We wish to preface this discussion with a general statement: *there is broad-based consensus in the field of evaluation that RCTs are an important methodological tool when applied under the correct circumstances*. However, if we are to apply RCTs correctly, we must have a rigorous understanding of what those circumstances are and when they are likely to be present.

We offer here some general comments followed by more specific comments organized by paper section.

General Comments

- **The paper needs to be more comprehensive and balanced.** The title of the paper ("What Constitutes Strong Evidence of a Program's Effectiveness?") leads the reader to expect a broad exploration of evidentiary strength and consideration of major relevant methods for achieving that strength. Yet internal validity is the only aspect of evidentiary strength that is seriously addressed and the RCT is the

only method seriously examined. Other methods are merely mentioned under Section II, or are used as foils to compare them unfavorably with RCTs.

We believe that the guidance needs to accomplish two objectives that are essential for OMB examiners and agency budget analysts: (1) explain what types of evaluations and methods are, and are not, acceptable and eligible to receive credit from OMB under PART questions 2.6 and 4.5; and (2) provide a framework for how to tell whether an evaluation submitted under the PART is technically sound and addresses the most important questions related to program performance. The document currently does not sufficiently achieve either of these objectives.

- **RCTs are weak with respect to the goal of program improvement.** RCTs have their greatest value in confirming or disconfirming specific hypotheses about whether or not a program is effective. RCTs are not very helpful in explaining *why* things happen and, except in the crudest sense (did the program work or not), do not contribute information for program improvement.

  Furthermore, RCTs do not provide information about *how* the program did or did not affect the outcomes in question. There is an important literature in evaluation on the rigorous modeling and study of program theory that was expressly developed to address this deficiency in RCTs. However, the paper does not mention this literature or suggest that studies of why and how the program affected the outcomes are essential in providing evidence of a program's effectiveness.

- **RCTs do not by themselves explicitly address construct validity.** In rigorous program effectiveness evaluation it is essential to assure that the program is implemented with high fidelity and that measures are reliable and valid. In technical terms, one needs to demonstrate the construct validity of the cause (i.e., the program) and the effect (i.e., the outcome measures). This is an empirical endeavor that should be carried out prior to mounting an RCT. The paper does not discuss the issue of construct validity in RCTs.

- **RCTs are weak with respect to generalizability or external validity.** Historically, one of the major criticisms of RCTs is that they are relatively weak with respect to external validity or generalizability. This critically important issue is not acknowledged in the paper, and, on the contrary, the reader is left with the impression that no such weakness exists. For example, under Section II, the paper states, "the difference in outcomes between the groups demonstrates the 'outcome' or impact one would expect for the intervention more generally" (p. 2). This is misleading. RCTs, if well done, assess outcome or impact at one point in time, in a particular setting or settings, at the current economic juncture, with the types of people who participated in the study, but surely not generally. In the Appendix, the guidance (under Quantity of Evidence Needed) calls for measures (multiple "typical" sites, etc.) that target some aspects of external validity but avoids addressing them as such.

The point here is not just that RCTs are strong with respect to internal validity but relatively weak for external. It is also that issues of external validity must be addressed because (a) using evaluation as an input to policy often requires generalizable information; and (b) the costs involved in dealing seriously with issues of generalizability might well invalidate what is written in Section IV, C, on the unlikelihood of prohibitive costs.

- **Addressing RCTs' validity problems often entails investment in companion program evaluations that have methodological designs other than RCTs.** This idea is well understood in biomedical research (where RCTs originated). In the clinical trials model in medicine, it is assumed that a potential intervention has already been through significant basic research that justifies the plausibility of the clinical intervention. Clinical trials involve four basic phases.

    - Phase I trials are exploratory small sample studies that examine tolerance of the treatment and potential side effects.
    - Phase II trials typically demonstrate whether the program is capable of achieving effects (efficacy) and examines its correlates.
    - Phase III trials are typically controlled effectiveness studies.
    - Phase IV trials typically examine generalizability and the fidelity of transfer of controlled interventions to field settings.

  Randomized designs are usually not used until late in phase II or more likely in phase III studies when effectiveness is the focus. The FDA reports that the vast majority of interventions (approximately 70-75%) that begin clinical trials do not survive to controlled Phase III randomized trials because they do not meet the basic conditions that warrant subsequent efforts. This critically important structure of clinical trials in medicine is not mentioned in the paper, despite the fact that the paper correctly cites this tradition as exemplary with respect to use of RCTs. OMB PART is already structured well to emulate this multi-phased model that situates RCTS as an essential component at the appropriate phase. By attending more carefully to the need for rigorous evaluation in connection with program development, measurement, and implementation as implied in PART Questions 1-3, the OMB would significantly enhance the quality, rigor, and potentially cost efficiency of effectiveness evaluations undertaken in connection with Question 4.

- **The importance of mixed methods.** There is a general consensus regarding the value of mixed methods in contemporary evaluation based on the recognition that all methods, however good, are fallible and have unique biases and weaknesses. The use of multiple methodologies, especially those that are less likely to share related biases, is recognized as one of the strongest mechanisms for identifying and controlling such biases. In addition, evaluation, unlike traditional basic science, is typically required to address multiple purposes, rather than simply the generation of new knowledge. Evaluations need to provide information about

what happened and why, and to suggest ways that programs might be improved. RCTs alone are a relatively limited mechanism for some of these purposes. However, when RCTs are appropriate, they often can also be augmented with other methods, either sequentially or synchronously, that can significantly help broaden both their rigor and utility. For example, high quality RCTs conducted for the purpose of program improvement should, in order to be considered rigorous for this purpose, be required to incorporate other methods such as interviewing, participant observation or theory-driven modeling in order to identify how and why the observed effects occurred.

- **The need to address feasibility and resource issues realistically**. In this guidance it is critically important to explicitly acknowledge feasibility considerations (including, but not limited to, cost and time) and provide agencies the flexibility for choosing to conduct evaluations that best leverage limited evaluation resources.

- **The need to address equity and human subjects concerns realistically.** Virtually no mention is made of the requirement of the RCT that people be denied a potentially beneficial intervention for the sake of the experimental design. While this position may be defensible ethically, and is often so defended in medical clinical trials, it is one of the most common concerns raised regarding RCTs and deserves mention in this context. It is important to note that the calculus for weighing the costs and risks to individuals against the potential gain to society from high-quality evaluation may differ from field to field. For instance, the U.S. Congress recently passed legislation to halt an RCT of the Upward Bound program apparently because they did not believe that denying access to that program for the sake of the controlled test of its effectiveness was a reasonable political or policy trade-off.

Specific Comments

Introduction

The paper begins by stating that "The revised PART guidance this year underscores the need for agencies to think about the most appropriate type of evaluation to demonstrate the effectiveness of their programs" (p. 1). One would expect that the document would review some of the key terms or phrases. For instance, what is an "appropriate" type of evaluation? What is meant by evidence of effectiveness? Perhaps more important, one expects that there would be a presentation of the variety of methods that provide useful information about program effectiveness, with some consideration of the strengths and weaknesses of each. Instead, the paper immediately "points to the randomized controlled trial (RCT) as an example of the best type of evaluation to demonstrate actual program impact" (p. 1). The author then inserts several caveat sentences:

> "Yet, RCTs are not suitable for every program and generally can be employed only under very specific circumstances. Therefore, agencies often will need to consider alternative evaluation methodologies. In addition, even where it is not possible to demonstrate impact, use of evaluation to assist in the management of programs is extremely important." (p. 1)

Several approaches are mentioned in passing (Experimental or Randomized Controlled Trials, Direct Controlled Trials, Quasi-experimental Studies, and Non-Experimental Studies (Direct or Indirect)), to be considered later.

The paper lists several questions that will be covered. But critically important questions that should be included are not raised. For instance, it makes sense that agency staff would want to know about things like:

- When is the right time to evaluate a program in terms of effectiveness? How frequently should the program be evaluated?
- What types of programs should be evaluated with the program evaluation methodology as described in this paper?
- How does this evaluation guidance apply to different types of programs like:
    - Intervention programs
    - R&D programs
    - Service programs
- How should program evaluation be adapted to large complex program systems that involve different types of subprograms or program components?
- How should agency staff make the decision about what type of evaluation is most appropriate for their program?
- How should resource issues be addressed?
- How should an agency develop the capacity to implement or manage evaluations of program effectiveness?

It seems that much more detail on these types of questions would enable OMB examiners and agency staff to make a more informed judgment about how to decide what is appropriate evaluation for purposes of PART and how to help ensure that rigorous high-quality evaluation will be carried out.

I. How is program evaluation addressed in PART?

The Role of Evaluation in Program Development and Implementation
This section points out that PART Questions 2.6 and 4.5 are relevant to evaluation and provides the full text for those questions. While the paper mentions that many of the questions in Section III of PART are relevant to evaluation, it does not consider them further. However, the paper does not recognize at all that evaluation plays a critically important role in Section I: Program Purpose and Design. Evaluation is essential in addressing such questions. For instance, if a program is not well designed and its implementation does not reflect its intent, the program does not have construct validity, an essential precondition for assessing the program's effectiveness. One approach to address such questions is evaluability assessment, a method developed in response to

evaluation problems related to program design, and one that has been used in the federal government for quite awhile.

The Role of Program Design and Implementation Evaluation in Impact Evaluation
Section II of PART on Strategic Planning is also integrally dependent on evaluation but receives no mention in this paper. Questions 2.1 – 2.4 relate to the specification of goals and measures, both foundational to conducting an evaluation. It is essential that program goals are well articulated and that measures are of high quality (reliability and validity) and there are well-established evaluation procedures for accomplishing these tasks. Every program effectiveness design, including RCTs, makes the assumption that these steps have been done well prior to undertaking the effectiveness study, and the quality of effectiveness studies depends on this.

Section III of PART on Program Management is at the heart of evaluation and involves the measurement and monitoring of program performance. It makes little sense to undertake a high quality assessment of program effectiveness (Section IV) without assuring first that the program is functioning well. In fact, one might even argue that it would be irresponsible both fiscally and methodologically to initiate an effectiveness study for programs that cannot demonstrate that they have well defined models, goals, measures and procedures for managing implementation. Yet, no mention is made in this paper that a precondition of doing the effectiveness evaluations of Section IV (RCT or otherwise) should be prior evidence of good performance on these foundational areas.

The Role of Internal and External Stakeholders in Evaluation
Question 2.6, which is a focus of this paper states:

> 2.6: Are independent evaluations of sufficient scope and quality conducted on a regular basis or as needed to support program improvements and evaluate effectiveness and relevance to the problem, interest, or need?

There is a fundamental problem that is raised by the use of the term "independent" in this question. Independent is defined in the guidance as:

> To be independent, non-biased parties with no conflict of interest should conduct the evaluation. Evaluations conducted by the program itself should generally not be considered "independent;" (p. 31)

The primary problem is that this does not acknowledge the critically important role of all stakeholders in developing the evaluation. For instance, many of the questions in Sections I-III of the PART relate to the articulation of the model, goals and measures of the program and of its management when being implemented. High-quality evaluation practice and common sense require that the "internal" program staff be an integral part of these evaluation activities. In many organizations, program development and planning are conducted largely internally and involve a number of evaluation activities. Rigorous high-quality assessment of effectiveness depends on their being done well using the best evaluation methodologies. But Question 2.6 mentions only "independent" or external evaluation. While external evaluators may play a role in helping to facilitate the

articulation of program models, goals and measures and in helping set up systems for monitoring implementation, it is inconceivable that these tasks could be done well only through independent external evaluators. Yet these activities are essential for addressing the requirement of Question 2.6 that the evaluation address "support program improvements and evaluate effectiveness and relevance to the problem, interest, or need." Some guidance for agency staff on these issues would be critically important.

Summary
The paper should describe the range of evaluation methods and approaches that can be used to address questions other than 2.6 and 4.5, should consider their relative strengths and weaknesses in different settings, and should discuss the fact that high quality and rigorous effectiveness evaluation is predicated on doing these tasks well. There are many well-established evaluation methods that would include (but not be limited to): needs assessment, program logic modeling, performance measurement, interviewing, focus groups, survey development and qualitative methods, measurement systems development, the assessment of the quality of measurement (reliability and construct validity), implementation assessment, process evaluation, and evaluability assessment, to name a few. Each of these approaches has a considerable well-developed literature and would be standard fare in even an introductory evaluation course. It is generally suggested in the evaluation field that high quality rigorous evaluation requires that the issues in Sections I-III of the PART guidance be accomplished well and that programs that have not addressed these issues cannot be considered ready for outcome or effectiveness evaluation.

II. What are the most common ways to evaluate program performance?

This section begins with the statement "The most significant aspect of program effectiveness is *impact*—the outcome of the program, which otherwise would not have occurred without the program intervention" (p. 2). In one sense, this is true. All program activities are ultimately directed towards this end. However, one cannot reach impact without assuring first that the pathways to impact have been successfully navigated. Impact is, from this perspective, the end of a long chain of events and is dependent on assuring that none of these links is weak. This chain is well established in the PART structure and would include the work done on defining program purpose and design, program strategic planning and measurement, and program management and implementation. The evaluation of program design and implementation discussed in the previous section are an essential foundation to evaluations focused on impacts and effectiveness. This is foundational work, much of which also involves evaluation (e.g., program modeling, needs assessment, measurement development, implementation and process evaluation, etc.). In fact, it is likely that the vast majority of day-to-day evaluation work for most federal programs will need to be directed to the foundational evaluation efforts that must be in place before an effectiveness evaluation can be justified or warranted. No mention is made of this in the paper. Instead, the paper mentions in passing that "Where it is feasible to measure the impact of the program…" and then moves on to recommending RCTs. But agency staff and OMB examiners have no guidance about how these foundational evaluation requirements need to be addressed in

order for high-quality impact assessment to be feasible. For instance, how long does it typically take to achieve the conditions required for effectiveness evaluation?

Making Informed Choices about Design
The paper states unequivocally that "RCTs are generally the highest quality, unbiased evaluation to demonstrate the actual impact of the program" (p. 2). They then continue with

> However, these studies are not suitable or feasible for every program, and a variety of evaluation methods may need to be considered because Federal programs vary so dramatically. Other types of evaluations may provide useful information about the impact of a program (but should be scrutinized given the increased possibility of an erroneous conclusion) or can help address *how* or *why* a program is effective (or ineffective) (i.e., meeting performance targets, achieving efficiency, fulfilling stated purpose). (p. 2)

However, they do not present any guidance for how one decides whether an RCT is the most appropriate method and present no detailed accounting of the strengths and weaknesses of the many alternatives. It would seem that some detailed explication of methods and perhaps something like a decision tree would help both OMB examiners and agency staff to make this complex decision.

The paper briefly presents five broad classes of methods. However, the level of detail and the breadth of coverage do not provide the information necessary for guidance of OMB examiners or agency personnel who have not been trained in these methods.

Specific Issues Related to "Classes of Methods"
First, the authors present the RCT. While they mention that there are feasibility challenges with RCTs ("There are many programs for which it would not be possible to conduct an RCT" (p. 2)), they do not consider here any of the other potential problems or biases associated with these designs. However, since the remainder of the paper is essentially devoted to RCTs, we will address these issues in subsequent sections.

The second of these methods – Direct Controlled Trials – seems to be inaccurately labeled. For example, both RCTs and many quasi-experimental designs would be considered "controlled trials" and it is not clear how the term "direct" describes the distinction the paper is trying to make. They seem to be referring to the traditional notion of an "experiment" in physical sciences research, a design which does not utilize random assignment because such a mechanism is not needed when the phenomenon being studied is "well-behaved" and expectations of the non-experimental condition have been well established in previous research. Historically it would probably be most accurate to label these "experimental designs" as distinct from *randomized* experiments where the mechanism of random assignment is required for control of extraneous variables. They present no assessment of the strengths or weaknesses of this class of design, but the history of science suggests that they predate the randomized experiment and have played the foundational role in scientific research. Thus, it is puzzling that the paper does not list them first (since the ordering they choose seems to be from stronger to weaker in internal validity) or discuss the circumstances in which such designs might appropriately be selected for PART evaluation.

The presentation of the third type of method, the quasi-experiment, is incomplete and somewhat confusing. It neglects any mention of the prominent types of quasi-experiments (regression-discontinuity, nonequivalent groups, interrupted times series), several of which are strong alternatives to RCTs when internal validity is the priority and have several distinct advantages over RCTs in certain circumstances. The paper describes quasi-experiments as "comparison group studies", but this is a term that applies equally well to RCTs.

The fourth class of methods is "non-experimental direct analysis" and encompasses a wide variety of approaches including pre-post and longitudinal studies and "correlation analyses, surveys, questionnaires, participant observation studies, implementation studies, peer reviews, and case studies." This is a huge area of methodology that the paper immediately dismisses with the statement "These evaluations often lack rigor and may lead to false conclusions if used to measure program effectiveness, and therefore, should be used in limited situations and only when necessary." We find this dismissal inappropriate because it does not acknowledge that for each of these methods, there is an extensive literature on how to assure that they are accomplished with high quality and rigor. For example, the method of peer review is considered foundational to rigor and quality in scientific research, even in the assessment of the quality of RCTs in biomedical and health research. While peer review has its weaknesses, it remains a central method in the evaluation of scientific research projects. Similarly, the area of longitudinal research is well established both qualitatively and quantitatively and is capable of yielding causal inferences that rival the internal validity of RCTs. Yet the paper dismisses this whole area out of hand.

Finally, the paper presents non-experimental indirect analysis, described primarily as expert review. They almost rule out this approach entirely except when no alternative is available. Again, this fails to recognize: (a) that this approach can, when done well, be carried out with rigor and with the systematic use of empirical data; and (b) is a dominant approach in how the federal government and many other entities evaluate complex problems. For example, the hearings on the Challenger Space Shuttle disaster constituted a systematic expert panel review that included expert testimony and the consideration of data and was successful in identifying the cause of the disaster from an extremely complex set of variables. Many federal entities like the Institute of Medicine, the National Academies of Sciences, the National Research Council and the Government Accountability Office use such methods effectively to assess program effectiveness. Even in the area of clinical medicine that is often cited as exemplary by advocates of RCTs, the expert review panel increasingly constitutes a critically important evaluative mechanism that supplements for weaknesses in RCTs and meta-analysis, especially when a complex evidence base is being assessed. It is reasonable to expect that for many federal programs such approaches, if accomplished well, would provide an excellent and cost-effective mechanism for assessing program effectiveness and it is puzzling why it was simply ruled out in this context.

Value of Broad Consultation, Including for RCTs

The paper correctly advocates that agency staff should "consult with internal or external program evaluation experts, as appropriate, and OMB to identify other suitable evaluation methodologies to demonstrate a program's impact. Some sources of evaluation expertise may include the peer-reviewed literature for the relevant discipline, scientific organizations such as the National Academy of Sciences, think tanks, and research organizations" (p. 3). We agree that such consultation is an important factor in building the capacity of federal agencies to accomplish high quality effectiveness evaluation. However, the paper prefaces that statement with the phrase "When it is not possible to use RCTs to evaluate program impact…" suggesting that such consultation would not be warranted for RCTs. We disagree. The construction and implementation of high quality RCTs is a complex endeavor and federal agencies should be encouraged to seek outside professional assistance on such matters. It would be extremely helpful for agency staff if the paper provided clearer and more detailed guidelines on how appropriate consultants might best be identified for evaluation consultation.

Utility of References; the Need for More

The brief set of references that are provided to "assist in the decision of what type of evaluation will provide the most rigorous evidence appropriate and feasible, the PART guidance provides several links to references on program evaluation" are a good start but need to be annotated and augmented with other high quality citations. Many of these sources would be confusing to agency staff without training in evaluation or further orientation regarding their appropriateness and use under different circumstances.

III. What sorts of tests provide strong evidence of a program's effectiveness?

This section is essentially a consideration of how RCTs provide strong evidence of a program's effect. We agree that under certain circumstances RCTs indeed provide strong evidence of effectiveness. However, there is consensus in the field that in other circumstances other methods are more feasible or appropriate than RCTs and may also provide strong evidence of effectiveness. In several parts of this section, it is clear that the comments about the advantages are especially delimited to or focused on social programs rather than programs in general. It is important to recognize that the federal government addresses a wide array of programs other than social ones and that the comments in this section should be qualified appropriately for agencies in these other areas.

Conditions Under Which RCTs May Be Appropriate

A key statement in this paper and one that has led to considerable confusion is the following:

> Well-designed and implemented RCTs are considered the gold standard for evaluating an intervention's effectiveness across many diverse fields of human inquiry, such as medicine, welfare and employment, psychology, and education. (p. 4).

We believe that the labeling of the RCT as a "gold standard" is inaccurate in that it does not precisely describe the conditions under which the RCT would even by its proponents be considered the "strongest" design.

Recommended pre-requisites to undertaking an RCT often include:

- the program is well defined and has an articulated program model
- the program has been implemented consistently and with high fidelity
- there are high-quality (e.g., valid and reliable) outcome measures
- the program as implemented is capable of producing change
- there is sufficient statistical power to accomplish the study with high quality
- the participants can be kept unaware of the group (intervention or control) to which they have been assigned
- the random assignment can be implemented and maintained
- drop-out rates do not occur or do not differ by group
- the interest is in confirming whether the implemented program caused the observed outcomes
- there is no well-established empirical baseline or standard against which the program could be compared
- there is the potential for preexisting differences between the groups that would obscure the treatment effect
- there is little interest in whether the program would be effective generally in other settings, with other persons or in other time periods
- ethical and human subject protections have been approved and are in place
- it is morally and socially acceptable to deny the intervention to participants in order to test effectiveness

Put in other terms, it might be accurate to say that RCTs are the "gold standard" only when conditions like the above have been met. When such conditions are not present, RCTs may be susceptible to bias or may be unfeasible. Some of these conditions are described in the last section of the paper (pps. 12-13) and in the Appendix. But their omission here is important to the question of "What sorts of tests provide strong evidence of a program's effectiveness?" The "gold standard" terminology may misrepresent RCTs when presented without these critically important and limiting qualifiers.

Potential for Inappropriate Use of Evaluation Resources

We especially want to avoid the danger that well-intentioned OMB examiners might say to Agencies: "RCTs are the Gold Standard, therefore the RCT is all we are willing to accept." The real danger exists that RCTs will be undertaken in circumstances where they are neither appropriate nor feasible. This problem will in the end reflect poorly on the PART process and runs the danger of wasting precious evaluation resources and reducing the possibility that more appropriate and rigorous methods will be used to improve federal programs. Finally, the "gold standard" language inappropriately sets the RCT "above" other methods. In fact, most alternative methods when implemented with quality and rigor could be considered "gold standards" under the appropriate circumstances. There are situations where a quasi-experimental design, a survey instrument or an expert

peer review panel might legitimately be described as the "gold standard" for the work at hand. The more important issues for establishing standards of rigor and quality for program improvement are how to direct agency staff to the most appropriate evaluation method for the program and how to assure that the method is implemented with the highest quality and rigor feasible. The "gold standard" language does not enhance the task of improving the rigor of effectiveness evaluation.

Specific Issues
When describing the "unique advantage of random assignment" the paper states that "…assuming the trial is properly carried out (as described in the Appendix) – the resulting difference in outcomes between the intervention and control groups can confidently be attributed to the intervention and not to other factors" (p. 5). This is, of course, a big assumption. The design and implementation of experiments is a complex endeavor that requires considerable technical expertise that most federal agencies do not readily have. It cannot be assumed that without considerable effort and expertise this condition will be met.

The paper states "Properly designed, RCTs are the only method that can eliminate the risk of bias, which can adversely affect the results of the evaluation" (p. 5). This is simply incorrect. There are many types of biases in any evaluation project. RCTs are particularly designed to address one important type: selection bias, or the threat that prior differences between treated and control groups might affect outcomes.

In this section the paper argues that "'single group pre-post' study designs often produce erroneous results" (p. 5). While such designs are comparatively weak relative to RCTs with respect to internal validity, they can be strong designs where there is a well-established empirical database of longitudinal data that can act as an appropriate comparison standard. There are whole fields of study, most notably economics and epidemiology, where such approaches are frequently and legitimately used and RCTs would not be appropriate substitutes.

On the top of page 6 the paper discusses a few important design considerations for RCTs. But it ignores a host of other important design issues including (but not limited to): interaction effects; multiple treatment comparisons (e.g., in factorial designs); or controls for variability in data (e.g., ANACOVA or blocking). While these cannot be covered adequately in so introductory a paper, they are important in RCT design and omission of this level of complexity may create for agency staff the false impression that experimental design is more straightforward and feasible than may be the case in many applied contexts.

The paper describes investigations that underscore the limitations of comparison group studies relative to RCTs and state that "these investigations have shown that comparison-group studies in social policy (employment, training, welfare-to-work, education) often produce inaccurate estimates of an intervention's effects, because of unobservable differences between the intervention and comparison groups that differentially affect their outcomes" (p. 6). But the relatively few such investigations often compare well-designed

RCTs with poorly designed alternatives (as is the case in the example shown in the figure on page 7). Furthermore, the paper goes on to state that "Even when statistical techniques have been used to adjust for observed differences between the two groups, problems have been found" (p. 6). But problems have been found in all major evaluation methodologies, including RCTs (consider the literature in clinical medicine that shows how well-designed clinical trials led to results that have subsequently been overturned). The literature on statistical adjustments for selection bias is extensive and is a major focus in a number of fields including economics where the Nobel Prize was awarded for work on such adjustments.

IV. The application of Randomized Controlled Trials: where they are / are not possible

In general we find this discussion both useful and accurate. It helps to provide some critical considerations that are essential to deciding on whether RCTs are feasible, and includes a useful list of examples of RCTs in a wide variety of federal program evaluations. The section on costs of conducting RCTs does a nice job of distinguishing between the relative advantages of large multi-site trials and smaller, more cost-effective ones, and provides some indication of the length of time that RCTs might require. Subsection D in this section and the appendix provide some important caveats and considerations that should be taken into account when deciding upon and implementing RCTs.