

# American Evaluation Association

---

16 Sconticut Neck Road, #290 • Fairhaven, MA 02719 • www.eval.org • aea@kistcon.com • (888) 232-2275 • (508) 748-3326

**DATE:** 8/17/2010

**TO:** Mary K. Wakefield  
Administrator, Health Resources and Services Administration

**SUBJECT:** Comments on Evidence of Effectiveness of Maternal, Infant, and Early Childhood Home Visiting Programs

The American Evaluation Association (AEA) is pleased to submit comments on proposed criteria for evidence of effectiveness of home visiting program models for pregnant women, expectant fathers, and primary caregivers of children birth through kindergarten entry, in accordance with Federal Register Notice of 7-23-10.

AEA is a professional association of evaluators dedicated to the application and exploration of program evaluation, personnel evaluation, technology, and many other forms of evaluation. AEA has approximately 6000 members representing all 50 states and the District of Columbia as well as over 60 foreign countries.

In summary, AEA recognizes the importance of using “evidence based” models as a basis for distributing funds available under the Affordable Care Act’s Maternal, Infant, and Early Childhood Home Visiting Program. We believe that the proposed criteria and methodology for a systematic review of such models represent a thoughtful starting point for assessing the evidence of their effectiveness. However we do have concerns about how the studies upon which the evidence is based are rated. We offer recommendations to 1) forego assigning an automatic high rating for random assignment designs and automatically relegating all other evaluation designs to moderate or low ratings, and avoid using the label “gold standard” in connection with random assignment designs in the rating methodology, 2) use additional criteria to assess the value of impact evaluations, 3) more specifically identify alternative impact evaluation methods, and 4) emphasize the value of multiple studies and mixed methods.

We hope our attached comments are helpful. If we can be of assistance, or if you need more information on our comments, please do not hesitate to call on us or to contact George Grob, our senior advisor for evaluation policy ([GeorgeFGrob@cs.com](mailto:GeorgeFGrob@cs.com), 540-454-2888).

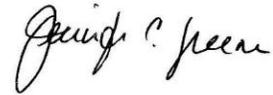
Sincerely,



Leslie Cooksy  
President



Debra J Rog  
Immediate Past President



Jennifer Greene  
President Elect

Attachment: Comments on Evidence of Effectiveness of Maternal, Infant, and Early Childhood Home Visiting Programs

# American Evaluation Association

---

16 Sconticut Neck Road, #290 • Fairhaven, MA 02719 • www.eval.org • aea@kistcon.com • (888) 232-2275 • (508) 748-3326

## **Comments on Evidence of Effectiveness of Maternal, Infant, and Early Childhood Home Visiting Programs**

The American Evaluation Association (AEA) is pleased to submit comments on the proposed criteria to be considered in assessing evidence of effectiveness of maternal, infant, and early childhood home visiting program models and on the methodology for a systematic review of such evidence, in accordance with Federal Register Notice of 7-23-10.

AEA is a professional association of evaluators dedicated to the application and exploration of program evaluation, personnel evaluation, technology, and many other forms of evaluation. AEA has approximately 6000 members representing all 50 states and the District of Columbia as well as over 60 foreign countries.

In summary, AEA recognizes the importance of using “evidence based” models as a basis for distributing funds available under the Affordable Care Act’s Maternal, Infant, and Early Childhood Home Visiting Program. We believe that the proposed criteria and methodology for a systematic review of such models represent a thoughtful starting point for assessing the evidence of their effectiveness. Our comments focus on the criteria for judging the effectiveness of the models. We offer recommendations to 1) forego assigning an automatic high rating for random assignment designs and automatically relegating all other evaluation designs to moderate or low ratings, and avoid using the label “gold standard” in connection with random assignment designs in the rating methodology, 2) use additional criteria to assess the value of impact evaluations, 3) more specifically identify alternative impact evaluation methods, and 4) emphasize the value of multiple studies and mixed methods.

### **AEA Evaluation Principles and Practices Regarding Evaluation Methodologies**

The AEA has sponsored numerous discussions and presentations on evaluation methods, including impact evaluations, through its annual conventions and in its professional journals, the peer reviewed *American Journal of Evaluation* and *New Directions for Evaluation*. In addition it has prepared “An Evaluation Roadmap for a More Effective Government,” a paper describing its vision of the role of evaluation in the Federal Government. The Roadmap outlines steps to strengthen the practice of evaluation throughout the life cycle of programs. It presents evaluation as an essential function of government that can enhance oversight and accountability of Federal programs, improve the effectiveness and efficiency of services, assess which programs are working and which are not, and provide critical information needed for making difficult decisions about them.

Particularly germane to our discussion here about the relative merits of impact evaluation methods and their application to the home health visitation program is the following excerpt from the Roadmap regarding analytic approaches and methods.

“Which analytic approaches and methods to use depends on the questions addressed, the kind of program evaluated, its implementation status, when the evaluation results are needed, what they are needed for, and the intended audience.

“No simple answers are available to questions about how well programs work, and no single analytic approach or method can decipher the inherent complexities in the program environment and assess the ultimate value of public programs. Furthermore, definitions of ‘success’ may be contested. A range of analytic methods is needed, and often several methods—including quantitative and qualitative approaches—should be used simultaneously. Some evaluation approaches are particularly helpful in a program’s early developmental stages, whereas others are more suited to ongoing and regularly implemented programs.”

These principles thus maintain that rigorous and useful evaluation is firmly anchored in the evaluation questions to be addressed and the developmental stage of the program to be evaluated. Programs with demonstrated efficacy and potential for broad implementation (among other preconditions for assessment) are best suited for rigorous impact evaluation.

It is in the context of these broad principles that we offer the following comments on the Department of Health and Human Services (HHS) proposal for assessing the value of studies that measure the impact of the Maternal, Infant, and Early Childhood Home Visitation Program.

## **Summary of Proposed Criteria for Assessing the Evidence of Home Visitation Models**

As a point of reference for our comments, we include here an excerpt from the Federal Register Notice that summarizes the rating scheme that the Department of Health and Human Services (HHS) proposes to use in deciding which home visitation models will be regarded as effective based on “well-designed, rigorous impact research” and therefore eligible for funding under the home visitation program.

“HHS proposes that an impact study will be considered high, moderate or low quality depending on the study’s capacity to provide unbiased estimates of program impact. Studies that are rated ‘high’ and ‘moderate’ (see Table 1 below) [not included here], therefore, would meet requirements to be considered ‘well-designed, rigorous impact research.’ In brief, the high rating would be reserved for random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment. The moderate rating would apply to studies that use a quasi-experimental design and to random assignment studies that, due to flaws in the study design or execution (for example, high sample attrition), do not meet all the criteria for the high rating. To receive the moderate rating, studies would have to demonstrate that at

the study's onset, the intervention and comparison groups were well matched on specified measures (i.e. baseline equivalence), such as a pretest measure of targeted outcomes or race and maternal education. Studies that do not meet all of the criteria for either high or moderate quality would be considered low quality.

“As summarized in Table 1 [not included here], the rating scheme would consider five dimensions: (1) Study design, (2) attrition, (3) baseline equivalence, (4) reassignment, and (5) confounding factors.”

In our comments below we may cite other sections of the Notice, but we believe that the excerpt above contains the key concepts that are germane to our comments, which now follow.

## **AEA Comments and Recommendations**

In summary, we believe that the HHS proposal for assessing the scientific rigor of evaluation designs and the resulting relative impact and value of home visitation models may result in over-rating the quality of some studies, under-rating some high quality evidence from quasi-experimental studies, and possibly mis-valuing the impact of some potentially effective home visitation models. We discuss our concerns and offer recommendations in the sections below.

### **I. Randomized Designs Are Not Necessarily Superior to Quasi-Experimental Designs in All Circumstances.**

**No Consensus on “Gold Standard.”** As stated in the Notice, HHS premises its preference for random assignment studies on its understanding that “Randomized control designs are often considered the ‘gold standard’ of research design because personal characteristics (before the program begins) do not affect whether someone is assigned to the program or control group.”

Within the professional evaluation community, there is no consensus on the relative weight of evidence accorded random assignment studies compared to other impact studies. In fact, this is a matter of some controversy. There is widespread agreement that randomized experiments can provide the most credible evidence of effectiveness under certain conditions. However, there are also limitations on what such studies can offer, and in other circumstances other methodologies may be better suited for measuring impact.

Recently, this topic has been the subject of three comprehensive reviews completed by the Government Accountability Office (GAO), the Congressional Research Service (CRS), and the Network of Networks for Impact Evaluation (NONIE) comprised of the Organization for Economic Co-operation and Development’s Development Assistance Committee (OECD/DAC) Evaluation Network, the United Nations Evaluation Group (UNEG), the Evaluation Cooperation Group (ECG), and the International Organization for Cooperation in Evaluation (IOCE).

For your convenience, the following citations and internet links are provided for easy reference:

- Program Evaluation: A Variety of Rigorous Methods Can Help Identify Effective Interventions, GAO (2009), <http://www.gao.gov/new.items/d1030.pdf>
- Congress and Program Evaluation: An Overview of Randomized Controlled Trials (RCTs) and Related Issues, CRS (2006), <http://digital.library.unt.edu/ark:/67531/metacrs9145/m1/>
- Impact Evaluations and Development, NONIE Guidance on Impact Evaluation, NONIE (2009), [http://siteresources.worldbank.org/EXT/OED/Resources/nonie\\_guidance.pdf](http://siteresources.worldbank.org/EXT/OED/Resources/nonie_guidance.pdf)

The last of these three studies was prepared in the context of international development. We recognize that HHS has legitimately excluded from its review impact studies of international development home visitation programs. Nevertheless, the NONIE report provides a more general perspective on multiple types of impact evaluations that are germane to U.S. domestic programs.

All three of these studies reached similar conclusions about random assignment studies. The following excerpt from the GAO report is typical of the conclusions in all three references.

“In our review of the literature on program evaluation methods, we found general agreement that well-conducted randomized experiments are best suited for assessing intervention effectiveness where multiple causal influences lead to uncertainty about what has caused observed results but, also, that they are often difficult to carry out. Randomized experiments are considered best suited for interventions in which exposure to the intervention can be controlled and the treatment and control groups’ experiences remain separate, intact, and distinct throughout the study. . . . Several other research designs are generally considered good alternatives to randomized experiments, especially when accompanied by specific features that help strengthen conclusions by ruling out plausible alternative explanations” (p. 20).

**Federal Practice in Assessing Effectiveness Also Varies.** In this regard, it is worth noting that as part of its review, GAO examined the practices of six Federally supported initiatives that identify effective interventions. While all of them consider the value of randomized studies, three of the six, all within HHS, (Evidence-Based Practice Centers at the Agency for Healthcare Research and Quality; Guide to Community Preventive Services at the Centers for Disease Control and Prevention; and National Registry of Evidence-Based Programs and Practices at the Substance Abuse and Mental Health Services Administration) do not require randomized experiments for interventions to receive their highest evidence rating.

**Potential Shortcomings of Randomized Experiments.** The rating system proposed for the home visitation program does attempt to deal with the limitations of random assignment studies by restricting its highest rating to those randomized studies with low attrition of sample members and no reassignment of sample members after the original random assignment. However, there are additional fundamental shortcomings that can limit the value of random assignment studies in providing evidence of the effectiveness of a program model, such as lack of fidelity to program design in the program’s implementation, inadequate construct validity of outcome measures,

limitations regarding the generalizability of the results, and dearth of information about the reasons for high or low impact. Here we highlight the last two of these shortcomings.

**Generalizability.** The argument for randomized experiments is that, if successfully conducted, they provide strong internal validity, that is, they provide a good test of whether the program as implemented made a difference in the outcome as measured. However, there may be a variety of reasons or conditions which limit the application of the program to settings other than those examined by the study.

For example, it may be that the program was not implemented well or that the outcomes were not well measured. Or, the program being evaluated and hence the entire impact evaluation may be narrowly limited to particular geographic, socioeconomic, and programmatic conditions that constrain the ability to predict whether the intervention would be successful under other circumstances.

**Explanatory Value.** In addition to methodological challenges in conducting randomized experiments, such studies provide little explanatory information regarding important causal mechanisms or contextual contingencies related to observed changes in valued outcomes. Random assignment studies may be able to demonstrate the impact of a program model, but are less likely to shed much light on the reasons for success or failure, or the opportunities to improve the intervention models.

With these and other kinds of limitations, when used alone, random assignment evaluations can fail to determine with confidence if the program had a strong impact on policy-relevant outcomes. Of course, problems of program implementation fidelity, valid outcome measurement, and generalizability of results are shared by other study methods as well. Nevertheless, these examples demonstrate the problems with focusing on only one type or aspect of validity when assigning priorities across methods.

All this having been said, we would not want to leave the impression that we underestimate the value of randomized assignment evaluations. Not only are there circumstances where they provide strong evidence of impact, but also the cumulative information they bring can be very helpful in meta-analytic studies. But as with all evaluative efforts applied outside the laboratory and in the real world, they have both strengths and weaknesses. For these reasons, we believe they should be rated on a broader set of criteria than are suggested in the HHS proposal.

**Recommendation 1.a. Forego assigning an automatic high rating for random assignment designs and automatically relegating all other evaluation designs to moderate or low ratings.**

In HHS's proposed rating scheme, a high rating would be reserved for random assignment studies with low attrition of sample members and no reassignment of sample members after the original random assignment. The proposed system risks two kinds of errors: rating as high quality a study with a randomized design that does not meet all the criteria for valid inference; or

rating as moderate or low what is actually high quality evidence from studies using quasi-experimental or other designs that have adequately addressed challenges to causal inference.

We therefore recommend that the high rating not be reserved for random assignment studies only, but that impact evaluations using other designs could also be eligible for a high rating if they, individually or in combination with other studies (possibly but not necessarily including random assignment studies) provide rigorous and credible evidence of effectiveness. Conversely, we recommend that those random assignment studies that fail to provide rigorous and credible evidence be rated as moderate or low, depending on a number of factors that we will discuss in this and other sections of our comments.

**Recommendation 1.b. Avoid using the label “gold standard” in connection with random assignment evaluation designs in the rating methodology.**

Given the lack of consensus in the professional evaluation community on the matter, it would be prudent to avoid using the label ‘gold standard’ which, as the HHS proposal notes, has sometimes been applied to random assignment studies alone. The continuing use of this designation can be harmful in that it may lead program sponsors or public agencies to commission expensive and time consuming randomized studies under conditions where other methods might be more appropriate. Furthermore, it may cause senior government officials and others to regard as unevaluated or unproven the value of effective programs which have been reviewed using well validated and contextually appropriate methodologies.

## **II. Many Other Factors, Especially Real World Conditions and Fidelity of Implementation, Affect the Quality of Scientific Evaluation Evidence**

The value of an impact study depends on more than its design type. It also depends on the details of the design, the manner of its implementation, and the results obtained, among other things. The HHS proposal recognizes these broader considerations by establishing a number of criteria to be used in assessing the relative effectiveness of home visitation models based on the quality of the studies it will review for that purpose. However, we believe that there are a number of additional criteria that might be useful to assess the quality of the studies. We offer the following standards as necessary for high-quality, rigorous impact assessment:

- adequacy of sample size to detect the effects that are expected
- the magnitude of the impact of the home visitation model found by the study
- validity and reliability of outcome measures
- implementation fidelity to program design
- appropriate correction for differences in baseline characteristics and nested effects
- appropriateness of data analyses and reporting
- clarity of description of the comparison group's experience
- the potential for or limitations to scaling up the program model or using it in circumstances that are different from those in the study
- evidence of the potential sustainability of results after the intervention has ended

- information from the study that helps explain the reasons for success or failure of the model being reviewed
- ethical considerations in replicating the model
- the extent to which the study rules out alternative explanations of success or failure of the model being reviewed
- the nature of, and process for selecting, study sample sites and participants
- assessment of the contextual factors of importance in program success or failure

For many of these criteria, there is no guarantee that a random assignment study will provide the kind of information that is most beneficial in deciding which program models are most effective. Other study types may be better suited in those cases.

**Recommendation 2. Provide additional criteria to use in assessing the value of impact evaluations.**

We believe additional criteria should be considered in rating the value of impact evaluations. Studies of poorly implemented models or narrowly designed studies likely warrant moderate or low ratings, despite having a random assignment design. Similarly, certain quasi-experimental approaches that have incorporated strategies for producing valid results may warrant the highest rating.

**III. Other Evaluation Designs Can Support Causal Attribution, Especially When They Can Rule Out Other Potential Causal Factors.**

The HHS proposal recognizes that quasi-experimental evaluation methods may produce rigorous evidence of effectiveness of home visitation program models. However, it does not identify or describe what these methods are. They are considered together in a category that accords them a value that is characterized simply as being inferior to randomized assignment studies in all evaluation contexts and for all home visitation programs. In fact, some of these methods are now highly developed and widely used by professional evaluators, many with strong results under appropriate circumstances. Quasi-experimental designs vary considerably in how well they can be expected, on average, to address the selection problem that is generally handled well by a random assignment experiment. However, they may be effective in dealing with selection bias in certain conditions or contexts. Furthermore, selection bias is not the only limitation of impact methodologies. The overall quality of the design depends on the extent to which, by the design and other study features, selection bias and other potential threats have been rendered implausible. All evaluation designs have various limits, including random assignment studies.

**Recommendation 3. More specifically identify alternative impact evaluation methods.**

We therefore recommend that the commonly accepted evidence based methods be named and described in the rating methodology, along with their relative advantages and the conditions under which they bring value in discerning evidence based impact of program models. The purpose of doing so would be to assist advocates of program models to determine whether their

models have a chance of being considered acceptable in the context of submitting proposals for funding under the home visitation program.

It is beyond the scope of these comments to construct the details of such a comparison, although they are described in the three references that we cited earlier. Here we would like to suggest that methodologies such as the following be specifically identified, described, and briefly assessed. Most of them are very different from the short examples in the HHS proposal which refers to quasi-experimental studies as involving self-selection or program selection by level of risk.

- **Purposefully chosen** comparison group designs, where individuals are selected to serve as a control group that resembles the treatment group as much as possible on variables related to the desired outcome
- **A pipeline approach**, which compares outcomes for the treatment group with households or individuals who have not yet experienced the treatment (delayed receipt), and notes the danger of contamination
- **Propensity scoring**, which statistically matches the treatment group with others with the same cluster of characteristics
- **Regression discontinuity design**, which compares outcomes for treatment and comparison groups that are formed by having scores above or below a cut-point on a quantitative eligibility or selection variable
- **Interrupted time-series design**, which compares trends in repeated measures of an outcome for a group before and after an intervention or policy is introduced

As noted earlier, each of these (and other) designs may be effective in certain conditions or contexts, perhaps with auxiliary methods or evidence. What counts here is not the study design per se, but rather the degree to which the design and other study features are successful in ruling out selection bias and other threats.

#### **IV. Knowledge of Program Impact is Enhanced by Considering Multiple Studies and Using Mixed Methods.**

Regardless of the perceived value of random assignment studies, relative to other approaches to impact evaluation, there is almost universal agreement within the evaluation profession that knowledge of program impact is enhanced by considering the results of multiple evaluations of programs and of evaluations performed using a combination of methods. In fact, independent replication of program outcomes under different conditions and with different methodologies can provide especially convincing evidence of program effectiveness.

<b>Recommendation 4. Emphasize the value of multiple studies and mixed methods.</b>
---

For this reason, we recommend that in assessing the evidence of the effectiveness of home visitation program models, the HHS review team consider all valid studies pertaining to the models it reviews. This could include in depth comparative case studies, contextually driven normative studies, process tracing, meta analyses, and various forms of syntheses of evidence, if conducted in tandem with or in addition to experimental or quasi-experimental studies. Ideally, conclusions about the relative effectiveness of various program models should be based on the body of work on that model and not just on a single study, if more than one valid study is available.

## **Conclusion**

In summary, AEA recognizes the importance of using “evidence based” models as a basis for distributing funds available under the Affordable Care Act’s Maternal, Infant, and Early Childhood Home Visiting Program. We believe that the proposed criteria and methodology for a systematic review of such models represent a thoughtful starting point for assessing the evidence of their effectiveness. However we do have concerns about how the studies upon which the evidence is based are rated. We offer recommendations to 1) forego assigning an automatic high rating for random assignment designs and automatically relegating all other evaluation designs to moderate or low ratings, and avoid using the label “gold standard” in connection with random assignment designs in the rating methodology, 2) use additional criteria to assess the value of impact evaluations, 3) more specifically identify alternative impact evaluation methods, and 4) emphasize the value of multiple studies and mixed methods.