How to Effectively Use the U.S. Department of Education's "Priority for Scientifically Based Evaluation Methods":

A Brief Guide for Program Offices

Coalition for Evidence-Based Policy
A Project Sponsored by The Council for Excellence in Government

WESTAT

March 2008

This publication was produced by the Coalition for Evidence-Based Policy, in partnership with Westat, under a contract with the Institute of Education Sciences (Contract # ED-04-CO-0059/0020). The Coalition is a nonprofit, nonpartisan organization sponsored by the Council for Excellence in Government (www.excelgov.org/evidence). The views expressed herein do not necessarily reflect the views of the Institute of Education Sciences.

We welcome comments and suggestions on this document (jbaron@excelgov.org).

Purpose of Guide and Overview of Key Recommendations

In 2005, the U.S. Department of Education announced a new policy tool – a "Priority for Scientifically Based Evaluation Methods" – that Department programs can use to encourage or require program awardees to have their projects rigorously evaluated (see *Federal Register* notice, attachment 1).

A. Purpose of this Guide: To advise program offices on using the "Priority" in an effective way, to expand the number of projects they fund that are rigorously evaluated.

B. Such evaluations can help serve key program objectives, such as:

1. **Building a body of research-proven strategies/models the program can then use to improve program performance.** In most program areas, interventions (i.e., strategies, models, practices) shown in rigorous evaluations to produce sizeable, sustained improvements in educational outcomes are rare or nonexistent. This leaves program officials and awardees with few research-proven tools that they can use to improve program performance. However, in some areas of education, such as early reading, dropout prevention, and substance abuse prevention, rigorous evaluations *have* successfully identified a few highly-effective interventions. Although rare, the very existence of these proven interventions suggests that a focused effort by Department programs to build the number of such interventions, and spur their widespread use, could produce major improvements in program effectiveness.

2. **Meeting the program's requirements for evaluation** – such as the evaluation requirements of the Government Performance and Results Act (GPRA) and the Office of Management and Budget's Program Assessment Rating Tool (PART).

C. A main challenge in using the Priority effectively: Ensuring that the resulting evaluations use a rigorous design and are conducted by a highly-capable evaluation team.

Many, perhaps most, attempts at rigorous evaluations of educational interventions fail to produce credible findings about the intervention's effectiveness because of serious flaws in study design or implementation. Illustrative examples of common flaws include: (i) failing to gain the cooperation of school officials and other stakeholders in the random assignment and/or data collection processes, resulting in non-equivalent control groups and/or incomplete data; (ii) using a sample too small to detect meaningful effects of the intervention; (iii) failing to obtain and analyze outcome data for a high proportion of the original sample; and (iv) measuring surrogate outcomes that lack practical and policy importance (e.g., attitudes toward school, rather than dropout and graduation rates).

Department programs that have used the Priority, and the experts they've engaged to review the proposed evaluations, have found it to be a valuable tool that has led to several high-quality evaluations. But, consistent with the experience above, they've also found it has generated too many proposals with flawed study designs and/or inexperienced evaluation teams, that would be unlikely to produce valid results. Based on these programs' experience, this Guide suggests a few concrete, streamlined procedures that program offices can use to help ensure the Priority attracts stronger proposals and leads to successful evaluations.

D. This Guide is organized into three sections:

(i) Brief description of the Priority;
(ii) Factors to consider in deciding *whether* to use the Priority; and
(iii) Once you go forward, suggestions for using the Priority in an effective way.

Overview of the Guide's Key Recommendations:

Factors to consider in deciding whether to use the Priority:

- As a threshold condition, the Priority is appropriate only for programs that (i) award funds competitively, and (ii) seek to improve one or more well-defined educational outcomes.
- As a favorable condition, at least some program awardees work with a sufficiently large number of individuals (e.g., students) or groups (e.g., schools) for a randomized controlled trial.
- Other favorable conditions for using the Priority in your program include the following:
  - If it is plausible to expect one or more projects you fund to improve educational outcomes within the time period of the project award.
  - If project outcomes can be measured using standardized tests, or other measures, whose ability to accurately assess outcomes is well-established.
- If you are still unsure whether it would be worthwhile for your program to use the Priority, we suggest you ask advice of an expert (e.g., staff of the Institute of Education Sciences).

Once you go forward, suggestions for using the Priority in an effective way:

- Enlist an expert in rigorous (especially randomized) evaluations to advise you on various aspects of the process.
- Decide whether to use the "absolute," "competitive preference," or "invitational" versions of the Priority, based on the degree to which you want to encourage rigorous evaluations.
- Provide program applicants and awardees, and their evaluators, with clear, practical guidance on conducting a rigorous evaluation (including the user-friendly guides noted in the main text).
- Ask your expert advisor, and an expert colleague that she recommends, to review the applicants' evaluation plans, and to suggest improvements where appropriate.
- Of particular importance, ask your expert reviewers to make sure the proposed evaluation team has a demonstrated track record in conducting the type of evaluation it is proposing.
- In the solicitation, ask applicants to verify that school officials and/or other key stakeholders support the evaluation, including random assignment where appropriate.
- Request periodic reports on each evaluation, once underway, to ensure it is adhering to the key items needed for success.
- Ask the awardee to submit a final report on the evaluation using the What Work Clearinghouse's *Guide To Reporting the Results of Your Study*.
- Ask your expert advisor to (i) review the reports to assess whether the studies produced valid evidence, and (ii) summarize the findings that are of greatest policy or practical importance.

I. Brief description of the "Priority for Scientifically Based Evaluation Methods"

The *Federal Register* notice containing the Priority is short and mostly self-explanatory (a copy is shown in attachment 1). What follows is a short summary of its key elements, and of what is required of program offices that seek to use the Priority.

A. Key elements of the *Federal Register* notice on the Priority:

1. The notice states that any appropriate Department program may use the Priority in a program solicitation, as an "absolute," "competitive preference," or "invitational" priority.

- If included in a solicitation as an "absolute" priority, the Priority requires all program awardees to have their projects rigorously evaluated using their award funds.

- If included as a "competitive preference" priority, it typically awards additional points (e.g., up to 25 in addition to the usual 100) to applications that include such a rigorous evaluation, with the number of additional points depending on the quality of the proposed evaluation.

- If included as an "invitational" priority, it indicates that the Department is interested in proposals that include such a rigorous evaluation, but will not give these proposals a competitive preference over other proposals.

2. The notice describes the rigorous evaluation methods that award applicants may use to qualify for the priority. When feasible, the project must use a randomized controlled trial. If random assignment is not feasible, the project may use either a matched comparison-group study or a regression discontinuity design. The notice states that, in general, under a competitive preference priority, randomized controlled trials receive more points than these other two designs.

The notice states that if grantees are focused on special populations in which sufficient numbers of participants are not available to support a randomized controlled trial or matched comparison-group study, single-subject designs that are capable of demonstrating causal relationships can be employed. (Such designs are sometimes required when program services are a matter of entitlement, as is often the case in the special education field.) However, there is little confidence that findings based on this design would be the same for other members of the population.

The notice includes short definitions of all of these study designs. It sometimes refers to randomized controlled trials as "experimental designs" and to the nonrandomized designs as "quasi-experimental designs."

3. The notice states that award applicants seeking to qualify for the Priority must include a plan for the rigorous evaluation. The plan must describe how the evaluator will collect valid and reliable data to measure the intervention's effect. The plan must also include: (i) the type of design the evaluation will use; (ii) the outcomes to be measured; (iii) a discussion of how the random assignment or matching will be done; and (iv) a proposed evaluator with the necessary expertise.

B. <u>What is required of a Department program seeking to use the Priority</u>:

1. **The program includes the attached *Federal Register* notice in its program solicitation, and states whether it is using the absolute, competitive preference, or invitational priority.** Factors to consider in deciding which form of the Priority to use are discussed in section III of this Guide (below). Program offices may download an electronic copy of the notice, to copy and paste into their solicitations, from http://www.ed.gov/legislation/FedRegister/finrule/2005-1/012505a.html.

2. **If used as a competitive preference priority, the program solicitation specifies how many additional points may be awarded to applications qualifying for the Priority** (e.g., 25 points in addition to the usual 100, depending on the quality of the proposed evaluation).

3. **The program solicitation requests an evaluation plan from applicants seeking to qualify for the Priority**. Typically, the solicitation would ask for a plan three pages in length, responding to the items listed in the *Federal Register* notice.

4. **If used as a competitive preference priority, the program office uses a two-stage process to review the applications, as described in the *Federal Register* notice.** In the first stage, the applications are reviewed without taking the Priority into account. In the second stage, the applications rated highest in stage one are reviewed for the competitive preference.

II. Factors to consider in deciding <u>whether</u> to use the Priority

A. <u>As a threshold condition</u>, the Priority is appropriate only for programs that (i) award funds competitively, and (ii) seek to improve one or more well-defined educational outcomes (e.g., student academic achievement, graduation rates, and postsecondary enrollment; abstinence from drug or alcohol use; teacher content knowledge).

Thus, programs that award funds by formula, as opposed to competition, would not be appropriate for the Priority because formula grant programs have little discretion over the terms by which funds are awarded. Also inappropriate for the Priority would be programs that award funds for a purpose other than improving a well-defined educational outcome, such as programs that fund the development of standardized tests to *measure* outcomes, or programs that provide funds to schools or universities to make infrastructure improvements such as building repair.

B. <u>As a favorable condition</u>, at least some program awardees work with a large enough number of individuals (e.g., students) or groups (e.g., schools) for a randomized controlled trial – which is cited in the Priority as the "best [method] for determining project effectiveness" and is thus most likely to help build a body of research-proven strategies and models. In order for an awardee to conduct a trial capable of determining whether their project is effective, the awardee must be able to randomly assign a sufficiently large sample to an intervention group that participates in the project, and to a control group that does not. Matched comparison-group studies (which the Priority cites as a second-best design) have sample size requirements similar to those of a randomized controlled trial. Regression discontinuity studies (also cited as a second-best design) have larger sample size requirements.

Programs whose awardees do not work with enough participants to support the above designs can still use the Priority to solicit evaluations using single-subject designs, but the ability of such designs to produce evidence that would generalize beyond those who participated in the study is less certain.

1. What follows are general rules of thumb on the sample size needed for a randomized controlled trial or matched comparison-group study in different types of programs.

These rules of thumb vary depending on whether the main goal of your program's evaluation strategy is to (i) identify only those projects with mid-to-large sized effects; or (ii) identify projects with small/modest effects too. Approach (ii) requires bigger samples, as discussed in the first endnote.[1] We present rules of thumb for both approaches.

- **Case #1: Your program funds awardees to develop and/or deliver interventions at the school or district level** – interventions such as schoolwide reform strategies, technical assistance in schoolwide reform, or assistance to districts in implementing data and accountability systems. In this case, a randomized (or matched comparison-group) evaluation of an awardee's project would likely need to allocate whole schools or districts – rather than individual students – to an intervention group and to a control or comparison group.

  ⇒ **In case #1, an awardee would need a sample of at least 10-20 schools or districts to identify a project with mid-to-large sized effects,** and would need a sample of at least 60-100 schools or districts to identify a project with small/modest effects.[2] (These numbers include both the intervention and control schools or districts.)

- **Case #2: Your program funds awardees to develop and/or deliver interventions at the teacher or classroom level** – interventions such as teacher professional development, or new classroom curricula. In this case, a randomized (or matched comparison-group) evaluation of an awardee's project would likely need to allocate teachers and classrooms to an intervention group and to a control or comparison group.

  ⇒ **In case #2, an awardee would need a sample of at least 20-30 teachers and classrooms to identify a project with mid-to-large sized effects**, and would need a sample of at least 100-150 teachers and classrooms to identify a project with small/modest effects.[3] (These numbers include both the intervention and control teachers and classrooms.)

- **Case #3: Your program funds awardees to develop and/or deliver interventions at the individual student level** – interventions such as dropout prevention or after-school interventions for at-risk students, scholarships, pre-doctoral or post-doctoral training, or individualized academic assistance such as tutoring. In this case, a randomized (or matched comparison-group) evaluation of an awardee's project would likely need to allocate individual students to an intervention group and to a control or comparison group.

  ⇒ **In case #3, an awardee would need a sample of at least 125-175 students to identify a project with mid-to-large sized effects**, and would need a sample of at least 700-1000 students to identify a project with small/modest effects.[4] (These numbers include both the intervention and control students.)

2. **Examples of programs <u>unlikely</u> to meet these sample-size requirements include the following:**

   - **Programs that provide financial assistance, such as loans or scholarships, directly to individual students.** Since each awardee is a single individual, the sample size is 1, so a randomized (or matched comparison-group) evaluation is not feasible. However, if the program – instead of funding students directly – makes grants to schools or universities to provide financial assistance to a sizeable number of students, some grantees may well have a sufficient sample of students to conduct such an evaluation.

   - **Programs that make grants to states to strengthen state-wide data and accountability systems.** Since each awardee is a state, and its data and accountability system affects all districts and schools in the state, the sample size is 1, so a randomized (or matched comparison-group) evaluation is not feasible. However, if the program makes grants to states to assist individual districts in implementing *district-level* accountability or other systems, some states may well have a sufficient sample of districts to conduct such an evaluation.

C. <u>Other favorable conditions</u> for using the Priority in your program include the following:

1. **If it is plausible to expect one or more projects you fund to improve educational outcomes within the time period of the project award.** For example, if one or more projects is designed to implement partly or fully-developed interventions (e.g., classroom curricula, teacher training models) that could reasonably be expected to improve educational outcomes within the time period of the project award (e.g., two to three years), that would be a factor weighing in favor of using the Priority. The ideal case would be if the projects could plausibly affect ultimate outcomes such as student achievement. But, short of that, conditions would still be favorable if the projects could plausibly affect nearer-term outcomes such as teacher knowledge or classroom practices.

   On the other hand, if all of the projects that you fund focus on early-stage development of interventions with little or no implementation in schools or classrooms, one would not expect the projects to improve educational outcomes during the course of the project award. In such as case, it would make little sense for your program to use the Priority to rigorously evaluate project outcomes.

2. **If project outcomes can be measured using standardized tests, or other measures, whose ability to accurately assess outcomes is well-established.** For example, if project outcomes can be readily measured using well-established achievement tests or objective, real-world measures (such as grade retentions, special education placements, graduation rates, attendance, and disciplinary suspensions), that would be a factor weighing in favor of using the Priority. The ideal case would be if project outcomes can readily be measuring using state test scores or other data (e.g., on attendance and graduation rates) that schools, districts, states, or universities *already collect for other purposes*. Using such "administrative" data eliminates the need for a project evaluation to administer its own tests, and may thereby reduce the cost of the evaluation substantially.
   At the other extreme, if measuring project outcomes would require the development and validation of new tests or other instruments, it may not be feasible to conduct project evaluations within your program at a reasonable cost. This factor would weigh against using the Priority.

D. <u>If you are still unsure</u>, based on the above factors, whether it would be worthwhile for your program to use the Priority, we suggest you ask advice of an expert in rigorous evaluations – such as staff of the Institute of Education Sciences, or another individual with the expertise described in the next section (IIIA, immediately below).

III. Once you go forward, suggestions for using the Priority in an effective way

What follows are a few suggested steps that program offices can take, with minimal administrative burden, to increase the Priority's likelihood of attracting strong proposals and producing successful evaluations.

A. Enlist an expert in rigorous (especially randomized) evaluations to advise you on various aspects of the process – such as the review of evaluation plans and final evaluation reports, as described below.

We suggest you select an individual with the following key qualification: *a solid understanding of the critical features that randomized controlled trials, matched comparison-group studies, and other study types cited in the Priority must have in order to produce valid evidence* – features such as those shown in attachment 2. The person need not necessarily be a researcher herself or intimately familiar with your program area. Her duties would be comprised of the activities described in C through I below.

To find such a person, you might start by asking staff from the Institute of Education Sciences to serve in this expert advisory role. If they are not available to do this, you could ask them to recommend other candidates within or outside the Department.

B. Decide whether to use the "absolute," "competitive preference," or "invitational" versions of the Priority, based on the degree to which you want to encourage rigorous evaluations. (These three versions of the Priority are described above, under section I.A.1, page 5.)

At one end, you might consider using the absolute priority (i.e., requirement) for rigorous evaluation if conditions in your program are highly favorable for use of the Priority, based on factors discussed in section II (above), or if your program has direction from Congress or the Department leadership to require rigorous evaluations. As an example, the Department's Striving Readers program has used the absolute priority based in part on strong Congressional support for rigorous evaluations in this program area.

A middle ground would be to use the competitive preference priority, in which you award additional points to applications that include a rigorous evaluation. Department programs that have used the competitive preference priority generally award a maximum of 20 or 25 points in addition to the usual 100. The strongest evaluation plans receive the maximum additional points; plans that are less strong but still meritorious receive fewer additional points.

At the other end, you might consider the invitational priority if for some reason a competitive preference priority is not feasible in your program. Although the invitational priority provides no meaningful incentive for award applicants to include a rigorous evaluation, it does signal applicants that a rigorous evaluation is a desirable use of award funds, and thus could encourage highly-motivated applicants to include such an evaluation.

C. Provide program applicants and awardees, and their evaluators, with clear, practical guidance on conducting a rigorous evaluation. For example:

1. Include, in the program solicitation, a link to the following resources --

- **The What Works Clearinghouse's *Key Items To Get Right When Conducting a Randomized Controlled Trial in Education*** (http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf). This is a brief guide for those seeking to undertake a randomized controlled trial. It describes, in plain language, key features that the trial must include in order to produce valid evidence about an intervention's effectiveness.

- **The What Works Clearinghouse's *How to Find A Capable Evaluator To Conduct a Rigorous Evaluation of an Educational Project or Practice: A Brief Guide*** (http://www.evidencebasedpolicy.org/docs/Guide-to-finding-evaluator-FINAL.pdf). This is a short, practical guide that program applicants (such as state and local educational agencies) may find useful in identifying a highly-capable evaluator to team with in order to qualify for the Priority. As discussed below, a highly-capable evaluator is usually critical to the study's success.

- **The Institute of Education Sciences' *Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide*** (http://www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf). This is a brief guide, intended primarily for those who are reading a study report on a completed randomized controlled trial or comparison-group study. It provides a concise overview, in plain language, of the key items to look for in assessing whether the study produced valid evidence about an intervention's effectiveness (excerpts from this guide are shown in attachment 2).

2. If your program has periodic meetings of applicants, awardees, and/or their evaluators, invite your expert advisor to speak to the group.

3. If resources permit, consider hiring an evaluation consultant to provide in-depth assistance to applicants, awardees, and/or evaluators in conducting rigorous evaluations. The Department's Striving Readers program has taken this approach in using the Priority. If you choose to hire such a consultant, we suggest you use the process described under E (below) to make sure that the consultant has a demonstrated track record in conducting rigorous – especially randomized – evaluations.

D. Ask your expert advisor, and an expert colleague that she recommends, to review the applicants' evaluation plans, and to suggest improvements where appropriate. We suggest two reviewers in order to bring two sources of expert judgment to the process. One way to identify the second expert is to ask your expert advisor to recommend a colleague with the key expertise in rigorous evaluation described above (under IIIA).

As described in the *Federal Register* notice, if your are using the competitive preference priority, your experts will only review, for the competitive preference, those applications rated highest in the first stage of the review process (where the Priority is not taken into account). Similarly, if you are using the invitational or absolute priorities, you could ask your experts to review the evaluation plans only for those applications that are serious contenders for award.

The experts will review the proposed evaluation according to the four criteria set out in the *Federal Register* notice (see boxed paragraphs in the notice at attachment 1). We suggest that you ask your experts not only to rate the quality of the applicants' evaluation plans, but to suggest improvements in the plans where appropriate. You can then ask the applicants to address the experts' suggestions before approving their funding award.

E. <u>Of particular importance</u>, ask your expert reviewers to make sure the proposed evaluation team has a demonstrated track record in conducting the type of evaluation it is proposing. We suggest you ask your experts to do this as part of their responsibility for assessing criterion 4 in the *Federal Register* notice, which asks them to make sure the evaluation plan includes "a proposed evaluator, preferably independent, with the necessary background and technical expertise to carry out the proposed evaluation."

- **<u>Why we emphasize this</u>: A demonstrated track record in rigorous evaluations is likely to be a stronger predictor of study success than anything the applicant might promise to do on paper.** Many attempts at a rigorous evaluation which appear promising in an application ultimately fail because the evaluation team does not have the capability to carry it out successfully. For example, the evaluation team may not have (i) the interpersonal skills needed to gain the cooperation of school officials and teachers needed to carry out the study, including the random assignment and data collection; (ii) the creativity to make revisions to the study design so as to address important needs (e.g., recruitment of an adequate sample) without compromising the design's validity (e.g., by violating random assignment); or (iii) the organizational skills to keep track of, and obtain outcome data from, a high proportion of the original sample members.

  We therefore strongly suggest asking your experts to verify, in assessing criterion 4, that the proposed evaluation team has the demonstrated ability to avoid such flaws and conduct a study that produces valid evidence.

- **To enable the experts to assess the evaluator's track record, ask each applicant to submit study reports on two evaluations their evaluator has conducted using the design they are proposing.** For example, if the evaluator is proposing a randomized controlled trial, ask for the study reports on two randomized controlled trials that the evaluator has previously conducted. It is probably not necessary that all members of the evaluation team have played a central role in these prior evaluations, just that one or two key team members did.

- **Then ask your experts to conduct a brief (e.g., 30-minute) review of each submitted study, to determine if it was well-designed and implemented.** You should make clear to your experts that you seek only a top-level review to assess whether the study was free of serious flaws in design and implementation. To facilitate the experts' review, we suggest you provide them with the excerpts from the Institute of Education Sciences' guide shown at attachment 2, which briefly lists key items to look for when reviewing a randomized controlled trial or matched comparison-group study. (Award applicants will also have access to this list if, as we suggest above, you include a link to the Institute of Education Sciences' guide in your program solicitation.)

F. In the solicitation, ask applicants to verify that school officials and/or other key stakeholders support the evaluation, including random assignment where appropriate. Such verification is important because the support of these stakeholders is usually a critical piece needed for the evaluation to go forward successfully. Verification might consist, for example, of a

brief letter from the superintendent of the school district in which the evaluation would take place, confirming his or her full support for the evaluation, including the random assignment and data collection. This letter would be included with the application.

G. Request periodic reports on each evaluation, once underway, to ensure it is adhering to the key items needed for success. Specifically, we suggest you ask each awardee that qualifies for the Priority, as a condition of its funding award, to provide brief quarterly or semi-annual updates on the evaluation's progress. For randomized controlled trials, we suggest you request that these updates address each item in the What Works Clearinghouse's *Key Items To Get Right When Conducting a Randomized Controlled Trial in Education* http://ies.ed.gov/ncee/wwc/pdf/guide_RCT.pdf). The key items described in this document are critical to the success of the study; it is therefore important that both you and the awardee remain vigilant throughout the course of the study to possible deviations from them.

We suggest you provide these periodic updates to your expert advisor, and ask her to identify any deviations from the *Key Items* that warrant concern. If she finds such deviations, we recommend that you bring them to the attention of the awardee as soon as possible, to enable corrective action.

H. Ask the awardee to submit a final report on the evaluation using the What Work Clearinghouse's *Guide To Reporting the Results of Your Study* – at http://ies.ed.gov/ncee/wwc/pdf/evaluator_guide.pdf. This guide provides clear, practical advice on reporting the results of an evaluation, so as to give the reader a clear understanding of what was evaluated, how it was evaluated, and what the evaluation found. Importantly, it asks the evaluation team to include a 1-2 page "structured abstract" that concisely summarizes the purpose of the study, setting, intervention, study sample, study design, and main findings.

I. Ask your expert advisor to (i) review the reports to assess whether the studies produced valid evidence, and (ii) summarize the findings that are of particular policy or practical importance. As part of the expert advisor's review, we suggest you ask her to use the excerpts from the Institute of Education Sciences' guide, shown at attachment 2, to help identify any flaws in the study that might undermine the validity of its findings. We suggest you also ask her to summarize findings from any of the studies that are of particular policy or practical significance – such as a finding that a particular project produced a sizeable effect on an important educational outcome, or a finding that a widely-used model produced little or no effect. You may wish to post the expert's summary on your program's website – along with a link to the actual study reports, if possible – in order to disseminate the evaluation results and enable others to learn from them.

Endnotes

[1] The reason a study needs a bigger sample to identify projects with small effects, as opposed to large effects, is as follows. The bigger the sample, the greater the confidence one can have that random assignment will result in an intervention and control group that are equivalent in key characteristics. Greater confidence that the two groups are equivalent provides greater confidence that any difference in outcomes between the two groups – even a small one – is due to the project and not to chance.

Although a bigger sample offers the important advantage of increasing the study's ability to detect small/modest effects, there may be tradeoffs to consider in seeking such samples from your awardees. First, some awardees may not be able to recruit a big sample, because they don't work with a sufficient number of schools, teachers, or students. Second, a bigger sample usually increases the cost of the study – for example, by increasing the data collection effort that is required.

The following endnotes are intended primarily for a research audience interested in the assumptions behind our sample size estimates on page 7.

[2] These estimates of sample size are based on the following assumptions: The desired power for the study is 0.80; the project's true effect size is at least 0.35 standard deviations (for a project with a mid to large-sized effect) or 0.15 (for a project with a small/modest effect); in each school or district 50 students participate in the study; the intra-class correlation is 0.07; a covariate (e.g., baseline test scores) with a 0.7 correlation with outcomes is used in estimating the project's effect; the study seeks to estimate the project's effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for all schools or districts in the original sample.

[3] These estimates of sample size are based on the following assumptions: The desired power for the study is 0.80; the project's true effect size is at least 0.35 standard deviations (for a project with a mid to large-sized effect) or 0.15 (for a project with a small/modest effect); in each classroom 25 students participate in the study; the intra-class correlation is 0.1; a covariate (e.g., baseline test scores) with a 0.7 correlation with outcomes is used in estimating the project's effect; the study seeks to estimate the project's effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for 90-100 percent of classrooms in the original sample.

[4] These estimates of sample size are based on the following assumptions: The desired power for the study is 0.80; the project's true effect size is at least 0.35 standard deviations (for a project with a mid to large-sized effect) or 0.15 (for a project with a small/modest effect); a covariate (e.g., baseline test scores) with a 0.7 correlation with outcomes is used in estimating the project's effect; the study seeks to estimate the project's effect at the .05 level of significance in a two-tailed test; and the study obtains outcome data for 80-90 percent of the original sample members.

ATTACHMENT 1 -- Federal Register notice on the Department's Priority for Scientifically Based Evaluation Methods

Federal Register

Tuesday,

January 25, 2005

Part II

Department of Education

Scientifically Based Evaluation Methods; Notice

DEPARTMENT OF EDUCATION

RIN 1890–ZA00

Scientifically Based Evaluation Methods

AGENCY: Department of Education.

ACTION: Notice of final priority.

SUMMARY: The Secretary of Education announces a priority that may be used for any appropriate programs in the Department of Education (Department) in FY 2005 and in later years. We take this action to focus Federal financial assistance on expanding the number of programs and projects Department-wide that are evaluated under rigorous scientifically based research methods in accordance with the Elementary and Secondary Education Act of 1965 (ESEA), as reauthorized by the No Child Left Behind Act of 2001 (NCLB). The definition of scientifically based research in section 9201(37) of NCLB includes other research designs in addition to the random assignment and quasi-experimental designs that are the subject of this priority. However, the Secretary considers random assignment and quasi-experimental designs to be the most rigorous methods to address the question of project effectiveness. While this action is of particular importance for programs authorized by NCLB, it is also an important tool for other programs and, for this reason, is being established for all Department programs. Establishing the priority on a Department-wide basis will permit any office to use the priority for a program for which it is appropriate.

EFFECTIVE DATE: This priority is effective February 24, 2005.

FOR FURTHER INFORMATION CONTACT: Margo K. Anderson, U.S. Department of Education, 400 Maryland Avenue, SW., room 4W333, Washington, DC 20202–5910. Telephone: (202) 205–3010.

If you use a telecommunications device for the deaf (TDD), you may call the Federal Relay Service (FRS) at 1–800–877–8339.

Individuals with disabilities may obtain this document in an alternative format (*e.g.*, Braille, large print, audiotape, or computer diskette) on request to the contact person listed under **FOR FURTHER INFORMATION CONTACT**.

SUPPLEMENTARY INFORMATION:

General

The ESEA as reauthorized by the NCLB uses the term *scientifically based research* more than 100 times in the context of evaluating programs to determine what works in education or ensuring that Federal funds are used to support activities and services that work. This final priority is intended to ensure that appropriate federally funded projects are evaluated using scientifically based research. Establishing this priority makes it possible for any office in the Department to encourage or to require appropriate projects to use scientifically based evaluation strategies to determine the effectiveness of a project intervention.

We published a notice of proposed priority in the **Federal Register** on November 4, 2003 (68 FR 62445). Except for a technical change to correct an error in the language of the priority, one minor clarifying change, and the addition of a definitions section, there are no differences between the notice of proposed priority and this notice of final priority. The definitions section provides the generally accepted meaning for technical terms used throughout the document.

Analysis of Comments

In response to our invitation in the notice of proposed priority, almost 300 parties submitted comments on the proposed priority. Although we received substantive comments, we determined that the comments did not warrant changes. However, we have reviewed the notice since its publication and have made a change based on that review. An analysis of the comments and changes is published as an appendix to this notice.

Note: This notice does not solicit applications. In any year in which we choose to use this priority, we invite applications for new awards under the applicable program through a notice in the **Federal Register**. When inviting applications we designate the priority as absolute, competitive preference, or invitational. The effect of each type of priority follows:

Absolute priority: Under an absolute priority we consider only applications that meet the priority (34 CFR 75.105(c)(3)).

Competitive preference priority: Under a competitive preference priority we give competitive preference to an application by either (1) awarding additional points, depending on how well or the extent to which the application meets the competitive preference priority (34 CFR 75.105(c)(2)(i)); or (2) selecting an application that meets the competitive priority over an application of comparable merit that does not meet the priority (34 CFR 75.105(c)(2)(ii)).

When using the priority to give competitive preference to an application, the Secretary will review applications using a two-stage process. In the first stage, the application will be reviewed without taking the priority into account. In the second stage of review, the applications rated highest in stage one will be reviewed for competitive preference.

Invitational priority: Under an invitational priority we are particularly interested in applications that meet the invitational priority. However, we do not give an application that meets the invitational priority a competitive or absolute preference over other applications (34 CFR 75.105(c)(1)).

Priority

The Secretary establishes a priority for projects proposing an evaluation plan that is based on rigorous scientifically based research methods to assess the effectiveness of a particular intervention. The Secretary intends that this priority will allow program participants and the Department to determine whether the project produces meaningful effects on student achievement or teacher performance.

Evaluation methods using an experimental design are best for determining project effectiveness. Thus, when feasible, the project must use an experimental design under which participants—*e.g.*, students, teachers, classrooms, or schools—are randomly assigned to participate in the project activities being evaluated or to a control group that does not participate in the project activities being evaluated.

If random assignment is not feasible, the project may use a quasi-experimental design with carefully matched comparison conditions. This alternative design attempts to approximate a randomly assigned control group by matching participants—*e.g.*, students, teachers, classrooms, or schools—with non-participants having similar pre-program characteristics.

In cases where random assignment is not possible and participation in the intervention is determined by a specified cutting point on a quantified continuum of scores, regression discontinuity designs may be employed.

For projects that are focused on special populations in which sufficient numbers of participants are not available to support random assignment or matched comparison group designs, single-subject designs such as multiple baseline or treatment-reversal or interrupted time series that are capable of demonstrating causal relationships can be employed.

Proposed evaluation strategies that use neither experimental designs with random assignment nor quasi-experimental designs using a matched comparison group nor regression discontinuity designs will not be considered responsive to the priority

when sufficient numbers of participants are available to support these designs. Evaluation strategies that involve too small a number of participants to support group designs must be capable of demonstrating the causal effects of an intervention or program on those participants.

The proposed evaluation plan must describe how the project evaluator will collect—before the project intervention commences and after it ends—valid and reliable data that measure the impact of participation in the program or in the comparison group.

If the priority is used as a competitive preference priority, points awarded under this priority will be determined by the quality of the proposed evaluation method. In determining the quality of the evaluation method, we will consider the extent to which the applicant presents a feasible, credible plan that includes the following:

(1) The type of design to be used (that is, random assignment or matched comparison). If matched comparison, include in the plan a discussion of why random assignment is not feasible.

(2) Outcomes to be measured.

(3) A discussion of how the applicant plans to assign students, teachers, classrooms, or schools to the project and control group or match them for comparison with other students, teachers, classrooms, or schools.

(4) A proposed evaluator, preferably independent, with the necessary background and technical expertise to carry out the proposed evaluation. An independent evaluator does not have any authority over the project and is not involved in its implementation.

In general, depending on the implemented program or project, under a competitive preference priority, random assignment evaluation methods will receive more points than matched comparison evaluation methods.

Definitions

As used in this notice—

Scientifically based research (section 9101(37) NCLB):

(A) Means research that involves the application of rigorous, systematic, and objective procedures to obtain reliable and valid knowledge relevant to education activities and programs; and

(B) Includes research that—

(i) Employs systematic, empirical methods that draw on observation or experiment;

(ii) Involves rigorous data analyses that are adequate to test the stated hypotheses and justify the general conclusions drawn;

(iii) Relies on measurements or observational methods that provide reliable and valid data across evaluators and observers, across multiple measurements and observations, and across studies by the same or different investigators;

(iv) Is evaluated using experimental or quasi-experimental designs in which individuals entities, programs, or activities are assigned to different conditions and with appropriate controls to evaluate the effects of the condition of interest, with a preference for random-assignment experiments, or other designs to the extent that those designs contain within-condition or across-condition controls;

(v) Ensures that experimental studies are presented in sufficient detail and clarity to allow for replication or, at a minimum, offer the opportunity to build systematically on their findings; and

(vi) Has been accepted by a peer-reviewed journal or approved by a panel of independent experts through a comparably rigorous, objective, and scientific review.

Random assignment or experimental design means random assignment of students, teachers, classrooms, or schools to participate in a project being evaluated (treatment group) or not participate in the project (control group). The effect of the project is the difference in outcomes between the treatment and control groups.

Quasi experimental designs include several designs that attempt to approximate a random assignment design.

Carefully matched comparison groups design means a quasi-experimental design in which project participants are matched with non-participants based on key characteristics that are thought to be related to the outcome.

Regression discontinuity design means a quasi-experimental design that closely approximates an experimental design. In a regression discontinuity design, participants are assigned to a treatment or control group based on a numerical rating or score of a variable unrelated to the treatment such as the rating of an application for funding. Eligible students, teachers, classrooms, or schools above a certain score (''cut score'') are assigned to the treatment group and those below the score are assigned to the control group. In the case of the scores of applicants' proposals for funding, the ''cut score'' is established at the point where the program funds available are exhausted.

Single subject design means a design that relies on the comparison of treatment effects on a single subject or group of single subjects. There is little confidence that findings based on this design would be the same for other members of the population.

Treatment reversal design means a single subject design in which a pre-treatment or baseline outcome measurement is compared with a post-treatment measure. Treatment would then be stopped for a period of time, a second baseline measure of the outcome would be taken, followed by a second application of the treatment or a different treatment. For example, this design might be used to evaluate a behavior modification program for disabled students with behavior disorders.

Multiple baseline design means a single subject design to address concerns about the effects of normal development, timing of the treatment, and amount of the treatment with treatment-reversal designs by using a varying time schedule for introduction of the treatment and/or treatments of different lengths or intensity.

Interrupted time series design means a quasi-experimental design in which the outcome of interest is measured multiple times before and after the treatment for program participants only.

Executive Order 12866

This notice of final priority has been reviewed in accordance with Executive Order 12866. Under the terms of the order, we have assessed the potential costs and benefits of this regulatory action.

The potential costs associated with the notice of final priority are those we have determined as necessary for administering applicable programs effectively and efficiently.

In assessing the potential costs and benefits—both quantitative and qualitative—of this notice of final priority, we have determined that the benefits of the final priority justify the costs.

We have also determined that this regulatory action does not unduly interfere with State, local, and tribal governments in the exercise of their governmental functions.

Intergovernmental Review

Some of the programs affected by this final priority are subject to Executive Order 12372 and the regulations in 34 CFR part 79. One of the objectives of the Executive order is to foster an intergovernmental partnership and a strengthened federalism. The Executive order relies on processes developed by State and local governments for coordination and review of proposed Federal financial assistance.

This document provides early notification of our specific plans and actions for these programs.

Electronic Access to This Document

You may view this document, as well as all other Department of Education documents published in the **Federal Register**, in text or Adobe Portable Document Format (PDF) on the Internet at the following site: *http://www.ed.gov/news/fedregister*.

To use PDF you must have Adobe Acrobat Reader, which is available free at this site. If you have questions about using PDF, call the U.S. Government Printing Office (GPO), toll free, at 1–888–293–6498; or in the Washington, DC, area at (202) 512–1530.

Note: The official version of this document is the document published in the **Federal Register**. Free Internet access to the official edition of the **Federal Register** and the Code of Federal Regulations is available on GPO Access at: *http://www.gpoaccess.gov/nara/index.html*.

(Catalog of Federal Domestic Assistance Number does not apply.)

Program Authority: ESEA, as reauthorized by the No Child Left Behind Act of 2001, Pub. L. 107–110, January 8, 2002.

Dated: January 17, 2005.

Rod Paige,

Secretary of Education.

ATTACHMENT 2 -- Excerpts from the Institute of Education Sciences' "Identifying and Implementing Educational Practices Supported By Rigorous Evidence: A User-Friendly Guide" (www.ed.gov/rschstat/research/pubs/rigorousevid/rigorousevid.pdf)

The first four pages summarize key items to look for when reviewing a randomized controlled trial to assess whether it produced valid evidence of an intervention's effectiveness. The last two pages summarize key items to look for when reviewing a comparison-group study. This is meant as a summary of general principles rather than an exhaustive list of features of well-designed studies.

Key items to look for in the study's description of the intervention and the random assignment process

1. The study should clearly describe (i) the intervention, including who administered it, who received it, and what it cost; (ii) how the intervention differed from what the control group received; and (iii) the logic of how the intervention is supposed to affect outcomes.

Example. A randomized controlled trial of a one-on-one tutoring program for beginning readers should discuss such items as:

- who conducted the tutoring (e.g., certified teachers, paraprofessionals, or undergraduate volunteers);
- what training they received in how to tutor;
- what curriculum they used to tutor, and other key features of the tutoring sessions (e.g., daily 20-minute sessions over a period of six-months);
- the age, reading achievement levels, and other relevant characteristics of the tutored students and controls;
- the cost of the tutoring intervention per student;
- the reading instruction received by the students in the control group (e.g., the school's pre-existing reading program); and
- the logic by which tutoring is supposed to improve reading outcomes.

2. Be alert to any indication that the random assignment process may have been compromised.

For example, did any individuals randomly assigned to the control group subsequently cross over to the intervention group? Or did individuals unhappy with their prospective assignment to either the intervention or control group have an opportunity to delay their entry into the study until another

opportunity arose for assignment to their preferred group? Such self-selection of individuals into their preferred groups undermines the random assignment process, and may well lead to inaccurate estimates of the intervention's effects.

Ideally, a study should describe the method of random assignment it used (e.g., coin toss or lottery), and what steps were taken to prevent undermining (e.g., asking an objective third party to administer the random assignment process). In reality, few studies – even well-designed trials – do this. But we recommend that you be alert to any indication that the random assignment process was compromised.

3. The study should provide data showing that there were no systematic differences between the intervention and control groups before the intervention.

As discussed above, the random assignment process ensures, to a high degree of confidence, that there are no systematic differences between the characteristics of the intervention and control groups prior to the intervention. However, in rare cases – particularly in smaller trials – random assignment might by chance produce intervention and control groups that differ systematically in various characteristics (e.g., academic achievement levels, socioeconomic status, ethnic mix). Such differences could lead to inaccurate results. Thus, the study should provide data showing that, before the intervention, the intervention and control groups did not differ systematically in the vast majority of measured characteristics (allowing that, by chance, there might have been some minor differences).

Key items to look for in the study's collection of outcome data

4. The study should use outcome measures that are "valid" – – i.e., that accurately measure the true outcomes that the intervention is designed to affect. Specifically:

- **To test academic achievement outcomes (e.g., reading/math skills), a study should use tests whose ability to accurately measure true skill levels is well-established** (for example, the Woodcock-Johnson Psychoeducational Battery, the Stanford Achievement Test, etc.).

- **Wherever possible, a study should use objective, "real-world" measures of the outcomes that the intervention is designed to affect** (e.g., for a delinquency prevention program, the students' official suspensions from school).

- **If outcomes are measured through interviews or observation, the interviewers/observers preferably should be kept unaware of who is in the intervention and control groups.**

  Such "blinding" of the interviewers/observers, where possible, helps protect against the possibility that any bias they may have (e.g., as proponents of the intervention) could influence their outcome measurements. Blinding would be appropriate, for example, in a study of a violence prevention program for elementary school students, where an outcome measure is the incidence of hitting on the playground as detected by an adult observer.

- **When study participants are asked to "self-report" outcomes, their reports should, if possible, be corroborated by independent and/or objective measures.**

  For instance, when participants in a substance-abuse or violence prevention program are asked to self-report their drug or tobacco use or criminal behavior, they tend to under-report such undesir-

able behaviors. In some cases, this may lead to inaccurate study results, depending on whether the intervention and control groups under-report by different amounts.

Thus, studies that use such self-reported outcomes should, if possible, corroborate them with other measures (e.g., saliva thiocyanate tests for smoking, official arrest data, third-party observations).

5. The percent of study participants that the study has lost track of when collecting outcome data should be small, and should not differ between the intervention and control groups.

A general guideline is that the study should lose track of fewer than 25 percent of the individuals originally randomized – the fewer lost, the better. This is sometimes referred to as the requirement for "low attrition." (Studies that choose to follow only a representative subsample of the randomized individuals should lose track of less than 25 percent of the subsample.)

Furthermore, the percentage of subjects lost track of should be approximately the same for the intervention and the control groups. This is because differential losses between the two groups can create systematic differences between the two groups, and thereby lead to inaccurate estimates of the intervention's effect. This is sometimes referred to as the requirement for "no differential attrition."

6. The study should collect and report outcome data even for those members of the intervention group who don't participate in or complete the intervention.

This is sometimes referred to as the study's use of an "intention-to-treat" approach, the importance of which is best illustrated with an example.

Example. Consider a randomized controlled trial of a school voucher program, in which students from disadvantaged backgrounds are randomly assigned to an intervention group – whose members are offered vouchers to attend private school – or to a control group that does not receive voucher offers. It's likely that some of the students in the intervention group will not accept their voucher offers and will choose instead to remain in their existing schools. Suppose that, as may well be the case, these students as a group are less motivated to succeed than their counterparts who accept the offer. If the trial then drops the students not accepting the offer from the intervention group, leaving the more motivated students, it would be create a systematic difference between the intervention and control groups – namely, motivation level. Thus the study may well over-estimate the voucher program's effect on educational success, erroneously attributing a superior outcome for the intervention group to the vouchers when in fact it was due to the difference in motivation.

Therefore, the study should collect outcome data for all of the individuals randomly assigned to the intervention group, *whether they participated in the intervention or not*, and should use all such data in estimating the intervention's effect. The study should also report on how many of the individuals assigned to the intervention group actually participated in the intervention.

7. The study should preferably obtain data on long-term outcomes of the intervention, so that you can judge whether the intervention's effects were sustained over time.

This is important because the effect of many interventions diminishes substantially within 2-3 years after the intervention ends. This has been demonstrated in randomized controlled trials in diverse areas such as early reading, school-based substance-abuse prevention, prevention of childhood

depression, and welfare-to-work and employment. In most cases, it is the longer-term effect, rather than the immediate effect, that is of greatest practical and policy significance.

Key items to look for in the study's reporting of results

8. If the study claims that the intervention improves one or more outcomes, it should report (i) the size of the effect, and (ii) statistical tests showing the effect is unlikely to be due to chance.

Specifically, the study should report the size of the difference in outcomes between the intervention and control groups. It should also report the results of tests showing the difference is "statistically significant" at conventional levels -- generally the .05 level. Such a finding means that there is only a 1 in 20 probability that the difference could have occurred by chance if the intervention's true effect is zero.

a. In order to obtain such a finding of statistically significant effects, a study usually needs to have a relatively large sample size.

Text omitted; for an updated discussion of sample size, see page 7 of the main text.

b. If the study randomizes groups (e.g., schools) rather than individuals, the sample size that the study uses in tests for statistical significance should be the number of groups rather than the number of individuals in those groups.

Occasionally, a study will erroneously use the number of individuals as its sample size, and thus generate false findings of statistical significance.

Example. If a study randomly assigns two schools to an intervention group and two schools to a control group, the sample size that the study should use in tests for statistical significance is just four, regardless of how many hundreds of students are in the schools. (And it is very unlikely that such a small study could obtain a finding of statistical significance.)

c. The study should preferably report the size of the intervention's effects in easily understandable, real-world terms (e.g., an improvement in reading skill by two grade levels, a 20 percent reduction in weekly use of illicit drugs, a 20 percent increase in high school graduation rates).

It is important for a study to report the size of the intervention's effects in this way, in addition to whether the effects are statistically significant, so that you (the reader) can judge their educational importance. For example, it is possible that a study with a large sample size could show effects that are statistically significant but so small that they have little practical or policy significance (e.g., a 2 point increase in SAT scores). Unfortunately, some studies report only whether the intervention's effects are statistically significant, and not their magnitude.

Some studies describe the size of the intervention's effects in "standardized effect sizes."[16] A full discussion of this concept is beyond the scope of this Guide. We merely comment that standardized effect sizes may not accurately convey the educational importance of an intervention, and, when used, should preferably be translated into understandable, real-world terms like those above.

9. A study's claim that the intervention's effect on a subgroup (e.g., Hispanic students) is different than its effect on the overall population in the study should be treated with caution.

Specifically, we recommend that you look for corroborating evidence of such subgroup effects in other studies before accepting them as valid.

This is because a study will sometimes show different effects for different subgroups just by chance, particularly when the researchers examine a large number of subgroups and/or the subgroups contain a small number of individuals. For example, even if an intervention's true effect is the same on all subgroups, we would expect a study's analysis of 20 subgroups to "demonstrate" a different effect on one of those subgroups just by chance (at conventional levels of statistical significance). Thus, studies that engage in a post-hoc search for different subgroup effects (as some do) will sometimes turn up spurious effects rather than legitimate ones.

Example. In a large randomized controlled trial of aspirin for the emergency treatment of heart attacks, aspirin was found to be highly effective, resulting in a 23 percent reduction in vascular deaths at the one-month follow-up. To illustrate the unreliability of subgroup analyses, these overall results were subdivided by the patients' astrological birth signs into 12 subgroups. Aspirin's effects were similar in most subgroups to those for the whole population. However, for two of the subgroups, Libra and Gemini, aspirin appeared to have no effect in reducing mortality. Clearly it would be wrong to conclude from this analysis that heart attack patients born under the astrological signs of Libra and Gemini do not benefit from aspirin. [17]

10. The study should report the intervention's effects on all the outcomes that the study measured, not just those for which there is a positive effect.

This is because if a study measures a large number of outcomes, it may, by chance alone, find positive (and statistically-significant) effects on one or a few of those outcomes. Thus, the study should report the intervention's effects on all measured outcomes so that you can judge whether the positive effects are the exception or the pattern.

Text that is not germane has been omitted.

A. Circumstances in which a comparison-group study can constitute "possible" evidence of effectiveness:

1. The study's intervention and comparison groups should be very closely matched in academic achievement levels, demographics, and other characteristics prior to the intervention.

The investigations, discussed in section I, that compare comparison-group designs with randomized controlled trials generally support the value of comparison-group designs in which the comparison group is *very closely matched* with the intervention group. In the context of education studies, the two groups should be matched closely in characteristics including:

- Prior test scores and other measures of academic achievement (preferably, the same measures that the study will use to evaluate outcomes for the two groups);
- Demographic characteristics, such as age, sex, ethnicity, poverty level, parents' educational attainment, and single or two-parent family background;
- Time period in which the two groups are studied (e.g., the two groups are children entering kindergarten in the same year as opposed to sequential years); and
- Methods used to collect outcome data (e.g., the same test of reading skills administered in the same way to both groups).

These investigations have also found that when the intervention and comparison groups differ in such characteristics, the study is unlikely to generate accurate results even when statistical techniques are then used to adjust for these differences in estimating the intervention's effects.

2. The comparison group should not be comprised of individuals who had the option to participate in the intervention but declined.

This is because individuals choosing not to participate in an intervention may differ systematically in their level of motivation and other important characteristics from the individuals who do choose to participate. The difference in motivation (or other characteristics) may itself lead to different outcomes for the two groups, and thus contaminate the study's estimates of the intervention's effects.

Therefore, the comparison group should be comprised of individuals who did not have the option to participate in the intervention, rather than individuals who had the option but declined.

3. The study should preferably choose the intervention/comparison groups and outcome measures "prospectively" – that is, *before* the intervention is administered.

This is because if the groups and outcomes measures are chosen by the researchers *after* the intervention is administered ("retrospectively"), the researchers may consciously or unconsciously select groups and outcome measures so as to generate their desired results. Furthermore, it is often difficult or impossible for the reader of the study to determine whether the researchers did so.

Prospective comparison-group studies are, like randomized controlled trials, much less susceptible to this problem. In the words of the director of drug evaluation for the Food and Drug Administration, "The great thing about a [randomized controlled trial or prospective comparison-group study] is that, within limits, you don't have to believe anybody or trust anybody. The planning for [the study] is prospective; they've written the protocol before they've done the study, and any deviation that you introduce later is completely visible." By contrast, in a retrospective study, "you always wonder how many ways they cut the data. It's very hard to be reassured, because there are no rules for doing it."[20]

4. The study should meet the guidelines set out in section II for a well-designed randomized controlled trial (other than guideline 2 concerning the random-assignment process).

That is, the study should use valid outcome measures, have low attrition, report tests for statistical significance, and so on.