



Ten Questions to Better Pilot Programs

Fiscal Brief

August 8, 2008

Changes needed in pilot program design to produce clear, useful results.

Executive Summary

The majority of pilot programs in North Carolina have failed to produce clear evidence of success or failure. This has made it difficult for members to determine whether or not to expand or discontinue the programs.

The General Assembly has expressed a strong desire to receive clear, objective evaluations of new programs. However, most pilot programs have been designed in ways that make quality evaluation impossible.

The goal of this memo is to help policymakers avoid the pitfalls that have undermined past pilot programs.

Policymakers should ask the following ten questions to ensure that new pilot programs will be able to provide clear results:

1. What is the problem that needs solving?
2. How does the program address the identified problem?
3. What is the cost of the program if it is successful?
4. Is there a budget or spending plan?
5. What criteria will be used to determine the program's success or failure?
6. What alternative programs/solutions might also address the problem?
7. Does the design of the program allow for meaningful evaluation?
8. Are there problems in the program design that will affect validity?
9. Is there sufficient time to observe effects?
10. Are there enough units of study to ensure statistical significance?

With clearer results, policymakers will better be able to determine which programs work and which programs do not.

Introduction

Pilot programs are new initiatives implemented on a limited basis as a test or trial. Ideally, the small-scale pilot program will provide data showing whether or not the new program has potential to succeed on a larger scale, or whether it should be discontinued.

The State of North Carolina has demonstrated an admirable willingness to try out new initiatives by funding new pilot programs. Unfortunately, policymakers have learned little from these efforts.

North Carolina's pilot programs have generally included provisions and funding for program assessment. Unfortunately, these pilot program assessments have often provided ambiguous results, making decisions on program expansion difficult for policymakers. The primary reason is that the pilot programs themselves have been designed in ways that inadvertently preclude meaningful assessment.

Common problems with pilot programs include:

- Unclear goals – what does it mean for a program to “work”?
- Unclear criteria – what measurements will be used to determine if a program is successful?
- No control group – results of the program are not compared against an independent group not affected by the pilot program;
- Selection bias problems – sites that are in the program are systematically different than those that are not;
- An inadequate timeframe in which to observe outcomes – some pilot programs have been discontinued before results can be observed;
- An inadequate number of pilot sites – the number of sites is insufficient to produce meaningful data.

The goal of this memo is to help policymakers avoid the pitfalls that have undermined past pilot programs.

By having program advocates answer the following ten questions, policymakers will ensure that new pilot programs funded by the State will be well-designed, and will provide clear results. With clearer results, policymakers will be able to determine which programs work and which programs do not, ensuring that taxpayer funds are directed to the best possible investments.

Question 1

What is the problem that needs solving?

It is important to first think about what problem or problems the State hopes to solve with a new pilot program. Developing a clear problem statement is the first and most crucial step in development of new pilot programs. That is, what is the nature, magnitude, and distribution of the social problem targeted by a program?

Developing the problem statement is important in that it provides a sense of direction to the project. Having a clear sense of direction is helpful for the project's evaluation and is also crucial during the implementation of the project. Success is more likely when those implementing the program have a clear sense of mission, understand the importance of the problem they are addressing, and have a clear understanding of what is expected of stakeholders.

Furthermore, pilot programs compete against other uses of State funds. Without understanding the problems that the pilot program aims to alleviate, it is very difficult to weigh competing claims on State resources.

When considering a problem statement, it is important to avoid problem statements that define the solution into the problem or contain causal claims.¹ Consider the statement "children are dropping out because of a lack of laptops in the classroom." This statement makes a causal claim (that drop-outs are a result of too few laptops) that might not be true, and defines the solution into the problem statement (provide more laptops). Instead, the problem statement should probably be "too many children are dropping out."

Question 2

How does this program address the identified problem?

Advocates for a new program or initiative should be able to clearly explain the theory or conceptual framework demonstrating why their new program or

initiative is the best way to solve an identified problem.ⁱⁱ There should be a clear, logical, and unambiguous relationship between the problem and the remedies that are to be applied to the problem.

Once the explanation of a program's theory or conceptual framework is given, the policymaker must critically assess the claims:

- *Do the program claims seem reasonable?* If something sounds too good to be true, it probably is. The vast majority of successful programs make improvements at the margin.
- *Is there any existing research backing the program's claims?* There is a chance that somebody has already studied the particular program being proposed. Members should consult with legislative staff to see what research exists.ⁱⁱⁱ
- *Are there any scenarios which could cause this proposal to fail?* It is important to consider the various ways a new program might fail, and ask whether the parties responsible for implementing the program have considered ways to avoid certain pitfalls. Additionally, the advocates might consider contingency plans if certain roadblocks to implementation should arise.

Question 3

What is the cost of the program if it is successful?

Pilot programs focus initially on a small target population or number of sites, in part to keep total costs manageable. It might be relatively easy for the State to find money to fund a pilot program. What happens though if the program is successful? Will the pilot program still be affordable if it is offered to the entire target population?

For example, consider a pilot program that costs \$1 million to deliver some new service to four schools. There are approximately 2,400 schools across the State, so expanding to a full-scale program statewide would cost \$600 million. However, costs would be substantially lower if the target population only included high poverty schools, or those schools tailored to students with special needs.

Members should consult with the Fiscal Research Division to determine how much a full-scale program would cost. If a full-scale program is something that would be cost-prohibitive, there is little point in conducting the pilot program.

Question 4

Is there a budget or spending plan?

Budgets can show policymakers whether or not the proposed pilot program:

- has been thoroughly planned;
- aligns spending to the program's stated goals; and
- includes the resources necessary for successful implementation and evaluation.

A well crafted, reasonable budget is an indication that thought has been given on how the new program will be executed. A vague, poorly crafted budget (or worse, a nonexistent budget!) is a likely indicator that the program has not undergone much of a planning process.

Additionally, a detailed budget allows policymakers to see if the spending plan aligns to the program's stated goals. Are the funds going towards services or products that address the identified problem, or will funds largely go towards staff salaries and overhead?

Finally, a detailed budget allows policymakers to see if the plan includes the resources necessary for successful implementation and evaluation. Two items, in particular, are commonly neglected in budgets for pilot programs:

1. professional development; and
2. program evaluation.

Consider a pilot program for a new drug treatment method. The health care workers implementing the program might require training to introduce the new program into practice. This expense might be significant, but could be crucial to successful implementation of the program.

Similarly, a program evaluation that provides reliable results can be expensive. As a result, this piece might be omitted from plans for new pilot programs. However, without proper evaluation, it is unlikely that the pilot program will generate unambiguous data.

One should consider that a quality budget plan is only an indication of how well the program is likely to be *implemented*. It is *not* necessarily an indication of how effective the program will be in achieving its objective. Even exceptionally well-run programs might not have a discernable impact on the problem.

Question 5

What criteria will be used to determine the program's success or failure?

Defining the performance criteria in advance increases the chances of getting clear results from a program. The criteria for evaluating a program should ideally be objective, unambiguous, and relevant to the program's goals.

Among education pilot programs, the typical criteria will be student achievement, as measured by student test scores. Other common criteria include dropout/graduation rates, and teacher turnover.

In addition to looking at overall test results, members might consider equity measurements. A new program could show great increases in test scores overall, but effects could vary widely between different groups of students. This might be a perfectly acceptable result. However it is important that policymakers demand that results are presented for various sub-categories of students, and – to the extent possible – define the levels of disparity that would be considered acceptable.

The criteria selection process might also identify data needs that can be worked out in advance of the initiation of the pilot program. The State agency in charge of implementation might be unable to accurately track certain measurements that program implementers would like to track. Prior to beginning a new program, it is important that parties implementing a new program work with the relevant State agencies to ensure that they can get the data needed to evaluate program outcomes. If the agency lacks the necessary capacity to gather the required data, policymakers must decide whether to provide the agency with the necessary capacity, or if they think the pilot program can proceed (still show meaningful results) using second-best data, or alternative criteria.

Ultimately, good pilot programs require evidence that can be reasonably obtained and serve as a credible basis for an answer. Without specific, measurable criteria, it will be difficult to tell if the evaluation question has been answered successfully or not.

Question 6

What other alternative programs/solutions might also address this problem?

For any identified problem, there are likely other programs, products or services being tried in other

states to address the problem. It is important that policymakers consider those and any other relevant alternatives before choosing to appropriate State funds. There might be alternatives that provide a greater likelihood of success, or can achieve similar ends at a lower cost.

Program advocates should be expected to disclose what other alternatives exist, when asked. Legislative staff are willing to research potential alternatives on a member's behalf.

Question 7

Does the design of the program allow for meaningful evaluation?

The most common reason pilot programs fail is that many are designed in such a way that evaluators are unable to demonstrate whether or not the pilot program is producing results. Ideally, the pilot program will show two things:

- 1) Members of the group to which the program is applied experience a specific outcome; and
- 2) Persons that are not members of the group do *not* experience that outcome.

Most valid studies fall into one of two design categories: a randomized controlled trial, or a comparison group study:

Randomized controlled trial: This design is one in which the units receiving the treatment are randomly assigned into either a treatment group (where units take part in the new pilot program) or a control group (where units do *not* take part in the pilot program). Researchers use randomized assignment to form two statistically equivalent groups in the most objective way possible.^{iv} This structure is the most statistically reliable technique to determine whether the observed outcomes are a product of the program, rather than a product of chance. Few, if any, pilot programs in North Carolina have included randomized controlled trials; however, most could easily have been designed as such with additional planning. Randomized controlled trials should be used whenever feasible.

Comparison-group study: In a comparison group study, there are still control groups and a treatment groups, however units are *not* randomly assigned to the groups. Instead, units are selected, to make the control and treatment groups as similar as

possible along an array of characteristics (e.g., demographics, educational achievement, etc.). With a comparison-group study, it is more difficult to demonstrate with confidence whether or not the pilot program has the intended effect. Careful attention must be paid to ensure that the two groups do not differ in any way that could affect the results. However such studies can be useful and might be the only option when implementation of a randomized controlled trial is not feasible.

There are many other study designs that do *not* allow for meaningful evaluation. These designs provide indications of *potential* program effects rather than conclusive findings. While not conclusive, these designs might help decide whether a more conclusive evaluation would be worthwhile.^v

“Pre-post” studies: This is a study that examines whether participants in an intervention improve or regress during the course of the intervention, and then attributes any such improvement or regression to the intervention.

Poorly-designed comparison-group studies: In many comparison-group studies the intervention and comparison groups are not closely matched.

Anecdotal evidence / satisfaction: Some program advocates present selected testimony or other measurements of participant satisfaction as evidence that a program is having the intended effect. Anecdotes do not rise to the level of evidence. When only anecdotal evidence is presented, it is likely a sign that the program lacks actual evidence of success.

Question 8

Are there problems in the program design that will affect validity?

Ideally, a new pilot program will produce results that have high validity. That is, the program will adequately demonstrate that:

- the intervention is actually causing the desired outcome (internal validity), and
- the program is replicable, producing similar results in different settings (external validity).

Randomized controlled trials inherently minimize most threats to validity. However, since few pilot programs in North Carolina are randomized controlled

trials, it is important that policymakers recognize some common threats to validity.^{vi}

Self-selection bias

One of the most common design flaws of North Carolina's pilot programs is self-selection bias. That is, pilot programs are conducted only in places that have expressed a desire to participate in the program. The problem is that participants' decisions to participate may be correlated with traits that affect the study. For example, schools that choose to participate in a pilot program might have teachers with higher levels of motivation than schools that choose not to participate. As a result, it may appear that the pilot program is working, when the results are really just a reflection of the quality of the participant schools' teachers. Such studies would fail to show that the intervention is actually causing the desired outcome.

Non-representative samples

Ideally, a pilot program tested on a small population will be deemed successful, and the full-scale program will be introduced to a larger population. Often, however, the units participating in the pilot program are not representative of the characteristics of the broader population that would be served under the full-scale program. For example, many pilot programs in North Carolina are introduced in the smallest counties, or the most economically disadvantaged areas. As a result, it is difficult to generalize the results. That is, will the program work across other districts in the state? For example, certain programs might be more effective in rural than urban areas, or the program might have differing effects on different minority groups. Policymakers will have a better idea if the pilot program is replicable if the treatment groups are as representative of the general population as possible.

Question 9

Is there sufficient time to observe effects?

Meaningful evaluation sometimes requires substantial time to observe a program's effects.

Some interventions might require time before they work. Educational programs that involve new ways of teaching, for example, might require a one or two-year ramp-up as teachers adapt to the new teaching method.

Other programs might be focused on long-term effects. In the case of a substance abuse program for young children, for example, it is the longer-term effect that

is of greatest significance. The evaluation of such a pilot must take the time to wait for long-term observations.

Additionally, substantial time is sometimes necessary to gather a sufficient number of observations. If a program shows an effect in year 1, it might simply be due to chance. But if the effects are replicated year-after-year, it is more likely that the observed effects are a result of the program intervention. With a greater number of observations, program evaluators are more likely to get meaningful results.

It is important to ensure that there is sufficient support behind a pilot program to allow enough time for evaluation to observe the desired effects. If the plug is likely to be pulled before the program produces results, it is best to refrain from even starting it.

Question 10

Are there enough units of study to ensure statistical significance?

In order for a study to be statistically significant, it needs to have a sufficiently large sample size. Larger sample sizes provide greater confidence that the outcome is a result of the pilot program intervention, rather than a result of chance. The required sample size varies based on what unit of study is chosen (students, classrooms, schools, or districts). The table below presents rules of thumb on sample sizes for educational pilot programs.^{vii}

Unit of study	Sample size (includes both control and intervention groups)
students	300
classrooms	50 - 60
schools	40 - 50
districts	15 - 20

The numbers in the table above are rules of thumb that will vary from study-to-study. Depending on the program, more or less units of study might be required. Fiscal Research analysts can work with parties designing new pilot programs to ensure that the program will include sufficient units of study to provide meaningful results.

Conclusion

These ten questions will guide policymakers toward better pilot programs. However, simply asking the questions is not enough.

Policymakers should insist upon pilot programs that are designed as randomly controlled trials whenever possible. A randomly controlled trial means that certain counties or districts will be receiving the pilot program intervention (the treatment group) while others will not (the control group).

Policymakers should avoid insisting that their district or districts be included in treatment groups. Being in the control group can be a good thing. First, not all pilot programs are helpful. Second, and more importantly, control groups are necessary to develop new programs that will eventually benefit *all* citizens across the State. A pilot program that generates actionable data is far more important than having a poorly designed program placed in a home district.

Additionally, policymakers should allow time for pilot programs to reach their full implementation and allow time to observe program effects. Acting too early might result in the abandonment of programs that are actually working.

A combination of smart policy design and a measure of political restraint is required for development of quality pilot programs. With better pilot programs, policymakers can make smarter investments in new programs and place North Carolina at the forefront of policy innovation.

For additional information, please contact:

*Kristopher Nordstrom
Fiscal Research Division
NC General Assembly
300 N. Salisbury St., Room 619
Raleigh, North Carolina 27603-5925
(919) 733-4910
<http://www.ncleg.net/fiscalresearch>
kristophern@ncleg.net*

ⁱ Bardach, Eugene, *A Practical Guide for Policy Analysis: The Eightfold Path to More Effective Problem Solving*. CQ Press (September 1, 2000), p. 5.

ⁱⁱ Program advocates are encouraged to develop a theory of change and logic model for their projects. Additional information on these topics can be found at the Centers for Disease Control website (<http://www.cdc.gov/eval/resources.htm#logic%20model>) and the W.K. Kellogg Foundation website (<http://www.wkkf.org/default.aspx?tabid=75&CID=281&NID=61&LanguageID=0>).

ⁱⁱⁱ Members desiring to do their own research may wish to begin their search at the What Works Clearinghouse (<http://ies.ed.gov/ncee/wwc/overview/>). The What Works Clearinghouse is a project of the U.S. Department of Education, and aims to provide education consumers with reviews of the effectiveness of various educational interventions.

^{iv} Myers, David and Mark Dynarski, "Random Assignment in Program Evaluation and Intervention Research: Questions and Answers," p. 2.

^v US Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance, "Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide," December 2003,

p. v.

^{vi} Information for this section based on lecture notes from Christina Gibson-Davis' Qualitative Evaluation Methods course (Pubpol 313) at Duke University, taken Spring 2005.

^{vii} US Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance, "Identifying and Implementing Educational Practices Supported by Rigorous Evidence: A User Friendly Guide," December 2003, p. 8.