Brief 2

Designing Better Pilot Programs: 10 Questions Policymakers Should Ask

Kristopher Nordstrom

Executive Summary

The majority of pilot programs in North Carolina have failed to produce clear evidence of success or failure, making it difficult for legislators to determine whether to expand or discontinue the programs. Although the General Assembly has expressed a strong desire to obtain clear, objective evaluations of new programs, the design of pilot programs often makes quality evaluation impossible.

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

To ensure that new pilot programs will provide clear, useful results, policymakers should answer the following 10 questions when reviewing program proposals:

1. What is the problem that needs to be solved?
2. How does the program address the identified problem?
3. What is the cost of taking the program to scale if it is successful?
4. Is there a budget or spending plan?
5. What outcome criteria will be used to determine the program's success or failure?
6. What alternative programs or solutions might also address the problem?
7. Does the design of the evaluation allow for meaningful results?
8. Are there problems in the evaluation design that will affect validity?
9. Is there sufficient time to observe effects?
10. Is the sample size large enough to identify statistically significant effects?

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

This brief was prepared by Kristopher Nordstrom as a fiscal brief from the NC General Assembly Fiscal Research Division, a staff agency of the NC General Assembly. Originally disseminated under the title *Ten Questions to Better Pilot Programs*, this brief was revised for the 2009 North Carolina Family Impact Seminar. Kristopher Nordstrom is a fiscal analyst with the Fiscal Research Division, North Carolina General Assembly.

Introduction

Pilot programs are new initiatives implemented on a limited basis as a test or trial. As part of any pilot, program implementers should collect sound data to show whether or not the new program has potential to succeed on a larger scale or whether it should be discontinued.

The North Carolina General Assembly has demonstrated an admirable willingness to try out new initiatives by funding pilot programs. North Carolina's pilot programs have generally included provisions and funding for program evaluation. Unfortunately, these evaluations have often provided ambiguous results, making decisions on program expansion difficult for policymakers. The primary reason is that the programs and/or their evaluations have been designed in ways that inadvertently preclude meaningful assessment.

Common problems with pilot programs include the following:

- 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 unaffected by the pilot program.
- Selection bias problems: sites that are in the program are systematically different from 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 brief is to help policymakers avoid the pitfalls that have undermined past pilot programs. Though this brief was written with pilot programs in mind, the recommendations could apply to the development of any new program.

When reviewing program proposals, policymakers should ask the following 10 questions. Answers should provide policymakers with the necessary information to assess program fit, feasibility, goals and evaluation design. Policymakers can then make informed funding decisions to ensure that new pilot programs are well designed and will provide unambiguous results. With clear 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 to be solved?

Developing a clear problem statement is the first and most crucial step in the development of new pilot programs. Both program developers and policymakers should be able to articulate the nature, magnitude and distribution of the social problem targeted by a potential new program.

Development of the problem statement provides program developers a sense of direction and guides both implementation and evaluation. Furthermore, with a thorough understanding of the problems that the program seeks to alleviate, policymakers will be better able to weigh funding choices against competing claims on state resources.

Program developers and policymakers should avoid problem statements that define the solution to the problem or contain causal claims.[1] For example, consider the statement "children are dropping out because of a lack of laptops in the classroom." This statement makes a causal claim (that dropping out is the result of too few laptops) that might not be true and defines the solution (provide more laptops). A more useful problem statement would 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 explain clearly the theory or conceptual framework that suggests the program will solve the identified problem.[2] 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 program's theory and conceptual framework are provided, the policymaker must critically assess the claims by asking the following:

- 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 existing research backing the program's claims? Legislators should consult with legislative staff to see if research exists.[3]
- Are there any scenarios that could cause this proposal to fail? Ask whether the parties responsible for implementing the program have considered possible pitfalls or roadblocks to implementation. What safeguards and contingency plans are in place?

Question 3: What is the cost of taking the program to scale if it is successful?

Pilot programs focus initially on a subset of the target population and a limited 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 will happen, however, if the program is successful? Will the program still be affordable if it is offered to the entire target population?

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

North Carolina legislators 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 unless the ultimate goal is limited replication.

Question 4: Is there a budget or spending plan?

Policymakers should examine budgets to assess 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 to how the new program will be executed. A vague, poorly crafted budget (or worse, no budget!) may indicate that the program has undergone only minimal planning. A detailed budget allows policymakers to assess whether the spending plan aligns with the program's stated goals (i.e., program priorities are well funded) and includes the resources necessary for successful implementation and evaluation. Two potentially critical planning and budget items are commonly neglected in pilot program budgets:

1. Professional development for program staff, and
2. Program evaluation.

Consider a pilot program focused on implementing a new drug-treatment method. The staff implementing the program might require training and careful supervision to introduce the new procedures into practice. The expense might be significant, but it could be crucial to the successful implementation of the program.

Similarly, a program evaluation that provides reliable results can be expensive. However, without proper evaluation, the pilot program will likely generate ambiguous data.

Bear in mind that a quality budget plan is an indication of how well the program is likely to be implemented, but 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 a problem.

Question 5: What outcome criteria will be used to determine the program's success or failure?

Policymakers should establish in advance the criteria for determining the success of a pilot program. What will a successful program accomplish? How will results be measured? How large does the program's effect need to be? The criteria for evaluating a program should be

objective, measurable, unambiguous and relevant to the program's goals.

For example, clear criteria for an education pilot program could include improvement in student test scores, dropout/graduation rates and teacher turnover. In addition to looking at overall test results, legislators might consider equity measurements. A new program could show great increases in test scores overall, but effects could vary widely among different groups of students. This might be a perfectly acceptable result. However, if the sample size is large enough, policymakers are encouraged to examine results for various subcategories of students and, to the extent possible, define the levels of disparity that would be considered acceptable.

Selecting the outcome criteria will allow policymakers to identify the types of data needed to evaluate the program. At the end of the evaluation, what specifically do policymakers want to know? Ideally, the pilot program will show two things:

1. Program participants experience a specific outcome; and
2. Similar persons that are not exposed to the program do not experience that outcome.

Program implementers should work with the relevant state agencies to ensure that they can get the data needed to evaluate program outcomes. If an agency lacks the capacity to gather the required data, policymakers must decide whether to provide the agency with the necessary capacity to gather it or to identify alternative criteria that will still show meaningful results.

Question 6: What alternative programs or solutions might also address the problem?

For any identified problem, there are likely 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 for a pilot program. There might be alternatives that provide a greater likelihood of success or can achieve similar ends at a lower cost. Program advocates should disclose what alternatives exist, when asked. Legislative staff can research potential alternatives on behalf of interested legislators.

Question 7: Does the design of the evaluation allow for meaningful results?

The most common reason pilot programs fail is that their evaluation designs do not allow evaluators to demonstrate the program's results clearly. Most rigorous evidence falls into one of two design categories: a randomized controlled trial or a comparison-group study with equivalent groups.

Randomized controlled trial: In this design, participants (e.g., individuals, schools, communities) are randomly assigned to either a treatment group (participation in the new pilot program) or a control group (no participation in the pilot program). Evaluators use random assignment to form two equivalent groups in the most objective way possible.[4] This structure is the most rigorous technique to determine whether the observed outcomes are a product of the program, rather than a product of other factors. Few pilot programs in North Carolina have included randomized controlled trials. However, most could 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 and treatment groups, but participants are not randomly assigned to the groups. Instead, participation in the two groups is based on observable characteristics (e.g., demographics) and evaluators strive to make the groups as similar as possible. With a comparison-group study, it is more difficult to demonstrate with confidence that an observed effect is caused by the pilot program. However, such studies can provide tentative support for a program 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 of a pilot program. These designs provide indications of potential program effects rather than conclusive findings.

They might help policymakers decide whether a more conclusive evaluation would be worthwhile.

"Pre-post" studies: Participants are assessed before and after the intervention. Changes are assumed to be caused by the intervention.

Poorly designed comparison-group studies: A comparison group is selected but is not closely matched with the treatment group on all relevant variables.

Anecdotal evidence / satisfaction: Selected testimony or a measurement of participant satisfaction is presented as evidence that a program is working. Anecdotes do not rise to the level of evidence. When only anecdotal evidence is presented, it is likely a sign that the program lacks meaningful evidence of success.[5]

Question 8: Are there problems in the evaluation 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 evaluated with randomized controlled trials, policymakers should examine evaluation results for some of the following common threats to validity.[6]

Self-selection bias: A common design flaw of North Carolina's pilot programs and of pilot programs generally 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 results. 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 differences in teacher motivation.

Non-representative samples: Pilot programs are tested with small samples of participants, and if deemed successful, they are scaled up to include a larger population. Often, however, the participants in the pilot program are not representative 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. Will the program work across other counties in the state? Certain programs might be more effective in rural than urban areas, or the program might have differing effects on different minority groups. Policymakers can be more confident that the pilot program is replicable if the sample participants are as representative as possible of the total population to be served.

Question 9: Is there sufficient time to observe effects?

Meaningful evaluation may require substantial time to observe a program's effects. 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, the evaluation must take time to wait for long-term observations of substance use.

Additionally, time is required to gather enough observations to determine if initial effects are replicated and maintained. If the observed effects are replicated year-after-year, it is more likely that they are a result of the program intervention. If individual outcomes last over time, they can be considered meaningful change.

Policymakers should ensure that a pilot program has sufficient support to allow time for meaningful evaluation. If the plug is likely to be pulled before the program is able to produce results, then it is not worth pursuing.

Question 10: Is the sample size large enough to identify statistically significant effects?

In order for study effects to be statistically significant, the study must have a sufficiently large sample size. The required sample size varies based on what unit of study is chosen (e.g., students, classrooms, schools, districts). The table below presents rules of thumb on sample sizes for educational pilot programs.[5]

Unit of Study	Sample Size (includes both control and intervention groups)
Students	300
Classrooms	50 - 60
Schools	40 - 50
Districts	15 - 20

Actual numbers required will vary from study to study. Depending on the program and outcomes assessed, more or fewer units of study might be required. Fiscal Research analysts can work with parties designing new pilot programs to ensure that the program will include a sufficient sample size to provide meaningful results.

Conclusion

Policymakers can use these 10 questions to guide their investments toward better pilot programs, which in turn will support the development of programs with better outcomes and better use of taxpayer dollars. However, simply asking the questions is not enough.

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

Policymakers should refrain from insisting that their districts be included in treatment groups. This is important to ensure that results are not skewed. Also, being in the control group can be beneficial. Not all pilot programs are helpful. More importantly, control groups are necessary to develop new programs that will eventually benefit all members of a target population. 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 demonstrate 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 the 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.

---

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

[2] 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 Web site (www.cdc.gov/eval/resources.htm#logic%20model) and the W.K. Kellogg Foundation Web site (www.wkkf.org/default.aspx?tabid=75&CID=281&NID=61&LanguageID=0).

[3] Members desiring to do their own research may wish to begin their search at the What Works Clearinghouse (ies.ed.gov/ncee/wwc/overview/) or the Promising Practices Network (www.promisingpractices.net/).

[4] Myers, D. & Dynarski, M. (2003). *Random Assignment in Program Evaluation and Intervention Research*: *Questions and Answers*. Washington, DC: Institute of Education Sciences.

[5] US Department of Education Institute of Education Sciences National Center for Education Evaluation and Regional Assistance. (2003). *Identifying and Implementing Educational Practices Supported by Rigorous Evidence*: *A User Friendly Guide*. Washington, DC.

[6] Information for this section is based on lecture notes from Christina Gibson-Davis's Qualitative Evaluation Methods course (Pubpol 313) at Duke University, taken Spring 2005.