The 2014 Farm Bill (Agriculture Act of 2014, H.R. 2642) includes $200 million for up to 10 three-year pilots in the Supplemental Nutrition Assistance Program (SNAP) “to reduce dependency and increase work requirements and work effort.” The Farm Bill also requires an evaluation of these initiatives.

This White Paper considers several issues related to the evaluation of the employment and training pilots: (i) random assignment; (ii) capturing entry effects; (iii) required sample sizes; (iv) outcomes and data collection strategies; and (v) the interrelation of the pilots and the evaluation.

The Legislation

SNAP (formerly known as the Food Stamp program) provides assistance to low-income households for the purchase of food. In total expenditure and in enrollees, SNAP is the country’s largest anti-poverty program, enrolling approximately 15 percent of the population in 2012.

While welfare reform in the mid-1990s imposed stringent work requirements within the Temporary Assistance for Needy Families (TANF) program, the requirements of the corresponding SNAP Employment and Training Program (“SNAP E&T”) are less rigorous and federal funding for the program is comparatively low. With some exceptions, states have the option of running voluntary and small programs. Most states do so.

In response to the rapid rise in the SNAP caseload and the relatively weak SNAP E&T program, the 2014 Farm Bill provided funds for pilot projects to improve our understanding of the likely impacts of various programmatic strategies on a range of outcomes including the caseload (i.e., dependency), employment, and earnings. Funding for SNAP E&T is likely to approximately double in states awarded pilot project grants.
The legislation specifically requires the U.S. Department of Agriculture (USDA) to select a diverse set of programmatic strategies. These might include: mandatory and strictly enforced job search requirements, child care subsidies, voluntary intensive job skills training, and voluntary, but intensive barrier removal (e.g., substance abuse and mental health treatment).

The SNAP E&T pilot programs are an exciting opportunity to learn about the likely impact of various programmatic strategies. The allocated funds are substantial; the discretion in choosing the programs to be implemented is broad. The balance of this White Paper considers the key decisions in structuring an evaluation of these pilot projects.

**Random Assignment**

The stated goal of the pilots is to learn about what works in SNAP E&T. To that end the legislation requires a “rigorous” evaluation of the “effectiveness” of the pilots and states are required to cooperate with that evaluation.

There are many approaches to estimating the impact of such pilot projects. Nevertheless, there is wide agreement that—when feasible—random assignment is clearly the preferred way to do so. With random assignment, some of those who are otherwise eligible are assigned to the pilot project and others are not; and the decision about who to assign to each group is made “randomly”—i.e., by the functional equivalent of a coin toss.

Such a random assignment design is strongly preferred because it mimics the thought experiment underlying the concept of impact: Otherwise identical people are assigned to the pilot project or not assigned to the pilot project. On average, random assignment holds all else equal, so the two groups differ only in that one group was offered the treatment, while the other was not. Thus, any difference in outcomes between the two groups can plausibly be attributed to the pilot project.

For a mandatory pilot project, the meaning of randomization is clear. Some people who receive SNAP benefits are required to participate
in the pilot project—and penalized if they do not; other people who receive SNAP are not required to participate in the pilot project. The evaluation then compares outcomes for those required to participate in the pilot project—whether or not they do participate—to outcomes for those not required to participate in the pilot project.

For a voluntary pilot project, the meaning of randomization is more subtle. For a voluntary pilot project, an evaluation will randomize who is offered the pilot project, perhaps among those deemed eligible. The evaluation then compares outcomes for those offered the pilot project—whether or not they actually participate in the pilot project—to outcomes for those not offered the pilot project.

Random assignment is clearly feasible for the SNAP E&T pilot projects; i.e., it should be possible to randomly split those eligible for a mandatory program or those applying for a voluntary program. That random assignment is feasible does not mean that it will be convenient or easy for program operators; there are certainly challenges to implementing random assignment. Random assignment is another step that must be inserted before service delivery can begin. Providers do not like to deny services. Volunteering beneficiaries do not like to be randomly denied services. Some mandatory beneficiaries will want to be in the group that avoids the requirement to participate. Finally, implementing random assignment and working with the evaluator requires moderate amounts of staff time. Skilled evaluators can work with program staff to minimize these concerns and inconveniences, but some time and inconvenience will remain.

Nevertheless, the goal of the legislation is to learn about the impacts of the pilot programs. The pilot programs are only funded in order for there to be programs to evaluate. No alternative to random assignment yields as strong evidence about program impacts. Compared to the alternatives, random assignment estimates are more precise, less subject to bias, and more convincing. Random assignment is critical for providing policy makers with the evidence needed to make future programmatic and funding decisions, and therefore clearly outweighs any inconvenience.
Capturing “Entry Effects”

Random assignment as described above applies to households receiving SNAP benefits. In a mandatory pilot project, some households eligible for the E&T program are assigned to the pilot project; the other households are not assigned to the pilot project. In a voluntary pilot project, some of households applying for the pilot project’s services are offered the services; the other households are not offered the services.

Such household-level random assignment estimates the impact on those who receive SNAP and are randomly assigned. However, it is plausible that the pilot projects—and the corresponding programs if fully rolled out—would affect who applies for/completes an application for SNAP. A mandatory and strictly enforced job search program might discourage some (perhaps many) households from completing applications for SNAP. Some observers have explained the sharp decline in the welfare caseload in the late-1990s as, in part, the effect of a mandatory job search program on suppressing entry onto welfare (rather than the effect of the job search assistance on exits from welfare). Conversely, an expensive voluntary program—for example, generous child care subsidies or expensive job training—that was only available to those on SNAP might encourage some additional households to apply for SNAP.

Conventional random assignment will not capture such entry effects because it only compares outcomes for those who receive SNAP and are randomized.

Instead, to capture such entry effects, an evaluation might consider geographic random assignment rather than household-level random assignment. In geographic random assignment, whether one was offered the pilot project’s services would depend on where one lives (e.g., the county of residence or the closest SNAP office). Some geographic areas would be randomly selected to get the services. In those geographic areas, everyone would receive or be offered the services. In the other geographic areas, no one would receive or be offered the services. Thus, in the geographic areas randomly assigned to the evaluation, applicants for SNAP could be certain that they
would receive or be offered the services. That certainty might affect their decision to apply for SNAP. Then, an evaluation could estimate entry effects by comparing outcomes between geographic areas with the pilot project and areas without the pilot project for everyone in the geographic area (e.g., the fraction of people receiving SNAP, average earnings).

**Required Sample Sizes**

Required sample sizes for useful evaluation results—and therefore required minimum pilot project size—are sensitive to exact assumptions. A useful rule of thumb is that for intensive interventions such as job training or a generous child care subsidy, samples of 1,000 households are often sufficient. However, for less intensive pilot projects, such as voluntary job search assistance (JSA), impacts are likely to be much smaller. As a result, to detect impacts much larger samples, perhaps 10,000 households, are needed. Sample size requirements for mandatory JSA pilot projects might be closer to those for intensive pilot projects than are those for voluntary JSA pilot projects.

Whatever the specific pilot project, it is important that there be approximate balance in the numbers of households randomized to being offered the pilot project and randomized to not being offered the pilot project. Designs in which almost everyone or every geographic area is offered the pilot project are unlikely to detect impacts.

In either case, larger samples will allow more precise estimates of which SNAP households would benefit more. Such information on differential impact would allow states to target the program to those who would benefit most, if funds were not sufficient to provide the program universally. Compared to the samples required to answer the question: “What works?” samples several times as large are required to answer the question: “What works best for whom?”

*For less intensive pilot projects, such as voluntary job search assistance (JSA), impacts are likely to be much smaller.*
Required sample sizes—i.e., number of SNAP cases randomized—to detect entry effects are much harder to estimate, but they are clearly larger than the sample sizes required to detect an impact for programs or services with no entry effects. Three or more times as many households randomized is often a useful rule of thumb (i.e., 3,000 or more for an intensive intervention and perhaps 30,000 or more for a low-cost intervention). Furthermore, those larger samples need to be in dozens of—randomly selected—geographic areas (e.g., counties, SNAP offices).

As a result, of the sample size requirements, detecting entry effects is likely to require something very different from a small pilot project. A state would need to identify dozens of geographic areas in which it could operate the pilot project, if that geographic area were randomly selected. Such a design may not be feasible in some of the smaller states or in states where the natural geographic areas do not lend themselves to random assignment (e.g., counties are too large or the SNAP caseload is overwhelmingly in one or a small number of undividable geographic areas).

**Outcomes and Data Collection Strategies**

The 2014 Farm Bill specifies interest in a wide-range of outcomes: the SNAP caseload, employment and earnings, and broader measures of well-being. Each of these outcomes is important. In particular, there is a plausible concern that mandatory job search assistance programs might push some people out of SNAP, without their finding employment—leading to lower total income and problems securing sufficient and nutritious food.

Crucially, the outcomes to be considered will affect the cost of the evaluation and plausibly its ability to detect impacts. The cost of an evaluation is usually driven by a combination of data collection strategy and sample size. For outcomes that are observed in administrative data for both those on and off of SNAP (e.g., SNAP caseload, SNAP benefits paid, TANF benefits paid, employment, earnings) evaluation cost is nearly independent of sample size. Service delivery costs will be approximately proportional to the number of people in the treatment group.
However, broader measures of household well-being (e.g., income, food security, household structure) are not measured in administrative data for the entire population. It is not enough to measure these outcomes for those on SNAP; some people will leave, perhaps because of the pilot project. Collecting such broader measures of household well-being would require a survey. Surveys of the SNAP population have incremental costs of several hundred dollars per household. Thus, surveying the several thousand households per pilot project who would need to be randomized to a low-intensity pilot project would drive evaluation costs, and may not be financially feasible.

This discussion suggests that the USDA and the states face important design trade-offs. Specifically, USDA and the states need to consider both the cost of the pilots and the cost of the evaluation. More intensive programmatic strategies have higher per participant costs, but larger likely impacts. From a pilot project cost perspective, states can choose either large low-cost pilot projects (i.e., job search assistance) or small higher-cost pilot projects (e.g., child care, job training, treatment for substance abuse or mental health). As long as the pilot project funds are not spread too thinly, either choice will probably yield pilot projects large enough to detect policy relevant impacts.

Given that these will often be new programs, it may be optimal to use the first year for refining the program models. Surveys and careful consideration of outcomes might start in the second year. It seems likely that samples would be large enough for this strategy.

The situation with respect to surveys is more subtle. The evaluation budget is probably not large enough to survey enough households to detect all but the largest plausible impacts of low-intensity pilot projects. Earnings is a fair—but far from ideal—proxy for broader measures of well-being. It would clearly be preferable to have survey-based measures of household income, food security, and...
other forms of hardship as well as subjective experiences with the pilot project and its non-compliance processes. It may not be cost feasible, however, to survey everyone who is randomized. For lower intensity, but much larger, pilot projects, and certainly for geographic randomization, only administrative data-based analyses may be feasible for the entire population.

**Interrelation of the Pilot Projects and the Evaluation**

The legislation gives USDA the dual responsibility of choosing the pilots from among state proposals and choosing the evaluator(s). USDA must release its plans for choosing the pilots by August 6, 2014, with applications due shortly thereafter and the awards announced in early February 2015. Presumably, the process for choosing the evaluator will follow a similar schedule. Given large dollar amounts, political attention, and these very short timelines, either of these tasks alone would be challenging.

However, USDA cannot deal with the two tasks—selecting the states and selecting the evaluator(s)—separately. Many details of the evaluation—whether to randomize, whether to randomize households
or geographic areas, at what point to randomize, how to randomize, how many households to randomize—need to be built into states’ pilot designs. Similarly, potential evaluators need to know whether and what form randomization will take.

This is a challenging problem and simply repeating in the RFA to the states the statutory language requiring states to cooperate with the evaluation is unlikely to be sufficient. In order to get the best possible evaluations, an attractive strategy would be to bring the evaluator on early enough to shape the RFA to the states. This might be done in three steps: (i) an initial letter of intent from states, sketching the pilot project concept and an approach to evaluation; (ii) comments from USDA or the evaluator about evaluability; and (iii) a second round proposal. The window for this strategy is closing quickly. An alternative would be to select the evaluator prior to awards to the state, allowing the evaluator to review state proposals for evaluability.

Short of such involvement of the evaluator in choosing the states, the best solution appears to be strong guidance requiring randomization—consistently provided to both states and evaluators. The need for strong guidance is particularly important if there is an interest in estimating entry effects. These pilot programs are among the most promising ever for estimating entry effects. Such entry effects are plausibly large and state SNAP E&T programs are potentially large enough and operating in enough different sites to make such entry effects detectable.

That said, estimating entry effects has rarely been done and would be challenging. If entry effects are of interest, USDA will need to choose states with respect to the evaluability of their proposed pilot projects. Specifically, USDA would need to specify geographic random assignment and require states to propose which geographic units would be randomized (i.e., in which it would offer the program, if randomly chosen) and approximate sample sizes in each geographic unit. Then, USDA needs to consider the feasibility of geographic random assignment.
Even with such strong guidance at the RFA stage, continued USDA involvement throughout the pilots is likely to be required. The requirements for rigorous evaluation are often challenging for program operators. The legislation requires states to cooperate with the evaluation. It is our experience that the best evaluations occur when agencies are willing to enforce the evaluation requirements expected by grantees. On the other side, USDA should insist that the evaluators make their evaluations as unobtrusive as possible.

This White Paper has made several suggestions about structuring the evaluations.

- Require random assignment. Random assignment is feasible and will yield the strongest possible evidence.

- If entry effects are of interest—and they probably should be for mandatory pilot projects—use geographic randomization. Some states may not be able to implement geographic randomization with sufficient scale to detect impacts.

- Select states that propose large samples—both of those offered the pilot project and of those randomized out of the pilot project. The larger the sample the more precise and useful the estimates.

- For intensive pilot projects (e.g., job training, generous child care subsidies, intensive barrier removal), sample sizes should be at least a thousand; ideally several times that.

- For less intensive pilot projects (e.g., job search assistance), sample sizes should be several times that size. It is possible, but not certain, that aggressive enforcement of job search requirements will have impacts similar to intensive pilot projects, and therefore require sample sizes in the low thousands.

- Carefully consider the outcomes of interest and signal the choices in the evaluation Request for Quotation. Unless the impacts are much larger than the previous literature would lead us to expect, it is probably not feasible to measure the impact of low-intensity broad pilot projects on survey-only outcomes, in particular for food security. Given the funds currently available for evaluation, trying to size a survey to do so will lead to under-funding other parts of the evaluation.

- Clearly signal key aspects of the evaluation—in particular, random assignment, perhaps geographic random assignment, and a strong preference for randomizing large numbers of SNAP households—in the RFA to the states and and the RFQ to the evaluator(s).
Implications for USDA, States and Evaluators

While the SNAP E&T pilots are an exciting opportunity, the challenges of evaluating such pilots are real. States need to establish and operate new programs, and evaluators need to execute large and complicated evaluations within very tight time frames.

Assuming that random assignment is used to evaluate these SNAP E&T pilot projects and there are no major obstacles, those evaluations should provide the nation with strong evidence about the likely impact of various programmatic strategies. Congress can then use that evidence about likely impact to reform the SNAP E&T Program to advance its goals.