National Association of Welfare Research and Statistics (NAWRS) /National Association of State TANF Administrators (NASTA) 2012 Research Academy

Summary Report

Mike Fishman
Jessica Wille

Table of Contents

Introduction

As part of their joint annual conference in September 2012, the National Association of Welfare Research and Statistics (NAWRS), together with the National Association of State Temporary Assistance for Needy Families (TANF) Administrators (NASTA), hosted a Research Academy focused on building state research capacity. With financial support from the Innovative Strategies for Increasing Self-Sufficiency (ISIS) project, the academy brought state and local administrators and researchers together with other research and evaluation experts to discuss how to incorporate rigorous research methods, particularly random assignment experiments, into ongoing state efforts to improve and document program effectiveness.

The Research Academy was envisioned by NAWRS and NASTA as a first step in a long-term agenda to build state research and evaluation capacity. The purpose of this summary report is to share highlights from the Research Academy in order to continue to build momentum for this important agenda. The report summarizes key points from the presentations at the Research Academy; presenter PowerPoint slides are attached.

The Research Academy consisted of an opening plenary session and three breakout sessions. The agenda was designed to:

- Present examples of the use of rigorous research methods, including random assignment experiments, to obtain information on program effectiveness with a focus on:
    - Experiments sponsored and/or conducted by state and local government;
    - Small-scale, quick turnaround evaluations using existing administrative data; and
    - Lessons from the private sector on the use of experiments to improve work processes.
- Engage participants in identifying current problems/issues that could be addressed by rigorous research, including the following:
    - How to turn an identified problem/issue into a "researchable question";
    - When and how a random assignment experiment could be used to answer that question; and
    - Whether data exist to capture key outcomes.
- Support skill-building among participants.
- Discuss barriers to state and local sponsorship of rigorous research and supports needed to address those barriers.
- Develop next steps for moving ahead locally and nationally on this agenda.

Plenary: Improving Program Effectiveness: A New Paradigm for Program Management

Mike Fishman, President of MEF Associates, chaired a panel that included Jim Riccio from MDRC, Steve Bell from Abt Associates, and Don Winstead from Don Winstead Consulting. Remarks from Howard Rolston, Abt Associates, were also shared as part of the plenary.

Setting the Stage for the Research Academy

Mike Fishman started off the plenary session and the Research Academy by calling for a new paradigm in the use of random assignment trials in social welfare programs in order to accelerate learning and program improvement.

He recounted David Brooks' April 26th *New York Times* column that featured Jim Manzi's recent book, "Uncontrolled," which makes the case for the expanded use of experiments or randomized control trials (RCTs) in order to learn what works in government programs. Manzi operates RCTs for private businesses and cites the thousands of experiments run by Citi and Google to identify effective business and marketing practices. Manzi concludes, and Brooks echoes his sentiment, that the government should sponsor more RCTs. Manzi calls for both creating a central government agency to oversee RCTs and decentralizing the use of RCTs.

Manzi takes some note of the use of RCTs for welfare research; however, he is not intimately familiar with how important RCTs have been to the shaping of welfare policy in the United States. Fishman noted that Howard Rolston's presentation (and his and Judy Gueron's upcoming book) outlines how experiments have played a fundamental role in the evolution of welfare policy over the past 40 years.

Fishman noted that we need to do more to accelerate the pace of learning. RCTs, as conducted today, are designed to answer big questions, often taking many years to produce findings. And they cost a lot of money – upwards of $20 million and more. Researchers test large service packages using a broad array of outcome measures. What is learned contributes to the policy and practice over time. However, the way RCTs are used does not contribute to rapid incremental learning.

Fishman wasn't suggesting an end to testing important programs that require a long-term perspective and expensive data collection regiments. Rather, he spoke to the need to implement experiments that help practitioners address more immediate problems, like figuring out how to get more welfare recipients to participate in work programs, how to run more effective job clubs, or how to better target case management services to make a difference in client behavior. By conducting small-scale RCTs using administrative data to test a panoply of practices and policies to see if they accomplish their desired ends, policy and program officials can test multiple variations at once to see which works better for whom and can manage their programs accordingly. The technology and know-how to do this are in place. A few states and local agencies have ventured out in this regard.

Fishman challenged the welfare administrators and researchers in the room to devote energy and resources to support the use of small-scale RCTs to promote learning and program effectiveness.

- Welfare administrators: Do not be satisfied managing for performance. Yes, performance is important, but you need to not only do things well, but to do the right things. And looking at outcomes only will be deceptive – you need impacts – and for many of your most important programs and policies you can only get reliable impacts from RCTs. Give your legislators

evidence that your programs are working and that you are working hard to make them work better.

- Welfare researchers: Work with state and local welfare administrators to help them identify opportunities for small-scale quick turnaround RCTs that can help them make incremental improvements in program operations. Give them a reason to rely on your expertise to help them better manage their programs.
- Federal program and policy officials: Look for ways to support state and local efforts to conduct their own RCTs. Consider putting 10–20% of what you invest into large RCTs to spur development in this area.

He noted this will accelerate learning at every level of the human services delivery system. The field can lower the stakes in the outcome of any one experiment by conducting hundreds of experiments that help make small modest improvements in how services are delivered that collectively can make a big difference in the lives of clients and the public's perception of the effectiveness of government.

The plenary panel spoke to this challenge.

- Howard Rolston (who Fishman spoke for as Rolston was unable to attend the conference) provided some background on the history of RCTs in the evolution of welfare policy.
- Jim Riccio spoke to his experience in both the United Kingdom (UK) and the United States (US) in using RCTs in a variety of settings and how that may transfer to conducting smaller-scale experiments at the state and local level.
- Steve Bell shared the work being sponsored by the Department of Education to promote the use of small-scale RCTs in schools around the country.
- Don Winstead shared his thoughts as a former state administer – both the opportunities and challenges facing states that would choose to move down this path.

A History of Random Assignment in Federal Programs

Mike Fishman presented on behalf of Howard Rolston. Rolston noted that the first large-scale social experiments were national/federal efforts that were designed to test policies in which these kinds of organizations (federal government or foundations) were interested.

The Income Maintenance experiments and the National Supported Work Demonstration had little state involvement in design or operation. In the late 1970s, the U.S. Department of Labor initiated the Work Incentive (WIN) Labs projects in which state WIN agencies developed and tested ideas that were evaluated with random assignment.

A big shift emerged in 1981–1982 with the passage of legislation that gave state welfare agencies greatly increased authority to run work programs for Aid to Families with Dependent Children (AFDC) recipients, including WIN Demonstrations, workfare, job search, and grant diversion. In retrospect, it is clear how this entirely changed the role of states in the evaluation of welfare reform efforts. Many state welfare agencies eagerly took up the new welfare-to-work options, but at the

federal level, there was initially little interest by the Reagan Administration in evaluating the effectiveness of these efforts. This opened the door for MDRC and the Ford Foundation to initiate a project to evaluate state and county programs that utilized the new work program authorities. And this brought states and counties into the forefront of experimental evaluations in that the conduct of the evaluations depended on welfare commissioners' and secretaries' willingness to subject their programs to a rigorous evaluation to determine if the programs worked. So what in the 1960s and 1970s was primarily a tool to evaluate demonstration strategies generated at the national level, now became a tool to evaluate programs that were designed and operated at the state and local level. Furthermore, these programs operated "in the wild" with large numbers of participants as opposed to being "hot house flowers" developed just for a demonstration.

Beginning in 1983, the U.S. Department of Health and Human Services (HHS) sponsored a series of demonstrations that use random assignment methodology, some of which overlapped with the MDRC/Ford Foundation effort. By this point, HHS and MDRC began working separately but cooperatively on these joint efforts. It is remarkable just how successful these first efforts were. Not only were most of the programs successful in increasing work and reducing welfare dependency, but they also had at least three other important outcomes:

- First, the findings were instrumental in leading to the Job Opportunities and Basic Skills Training program (JOBS) in 1988 and in increasing federal funding for work programs almost tenfold.
- Second, they established the feasibility of using random assignment in welfare offices to evaluate ongoing programs.
- Third, they relied on administrative records (Unemployment Insurance wage and welfare), allowing evaluations to operate at a much larger scale and at a lower cost.

In 1987, states began to request a significant number of waivers under section 1115 of the Social Security Act to test new reforms of AFDC. The 1115 waiver application required a plan for evaluating the effects of the proposed changes to the AFDC program. Given the proven feasibility and reliability of random assignment, HHS adopted the experimental method as the preferred evaluation approach. However, given considerable resistance by many (though not all) states, the preference was inconsistently enforced. But by 1992, when it was even more apparent that random assignment produced credible results and other methods did not, HHS began a policy of much stricter enforcement. Eventually, 45 states would receive waivers, with almost all using experimental designs. The evaluations were conducted under specifications jointly agreed to by the state and federal government, but the states themselves procured and managed them, expanding the capacity of state evaluation staff to run experiments.

While the waiver demonstrations were underway, HHS continued to partner with states on other evaluation efforts, such as the JOBS Evaluation (the name changed to National Evaluation of Welfare-to-Work Strategies, NEWWS, after the passage of the 1996 welfare reform law that ended the JOBS program). In addition, several states initiated major experimental evaluations, including California's GAIN (Greater Avenues for Independence) program and Florida's Project Independence. This gave states further experience with random assignment. These efforts

produced important findings. For example, evaluations found that work-first strategies were more effective in increasing work and reducing welfare and less expensive than remediation of basic education, that neither kind of program by itself increased participants' income, but that combining them with more generous treatment of earnings did. These findings were important in influencing the legislation that created TANF.

After the passage of TANF, partnerships between states and federal agencies on random assignment evaluations continued with the Employment Retention and Advancement (ERA) and the Hard-to Employ (HtE) demonstrations. Unfortunately, increasing retention and advancement among TANF recipients, and helping the hard-to-employ have proven difficult, and success in finding effective programs has been more sporadic than earlier.

Because of the partnership between the states and the federal government in conducting random assignment evaluations, there is much stronger evidence on how to move people from welfare to work than for creating success in other areas of social policy. But there is still much to be learned both for federal policy, and state and local policy and operations. For example, despite all that was learned about work-first programs, much less is known about what kinds of programs are effective at building soft and hard skills. But important experimental work in this area, in projects such as ISIS, is underway.

So there is much work to be done, both in federal efforts in which states are involved and state efforts on their own.

What Drives Evaluation Costs?

Jim Riccio next discussed what may be learned from the conduct of large-scale, big-question RCTs that is transferrable to smaller-scale studies. Riccio recounted that most of his experience is with large-scale, long-term, expensive RCTs. He used this experience to help illustrate what drives evaluation costs, and what considerations and choices can help an organization decide whether a small-scale RCT is feasible and sensible.

He used an example from the UK to add a little international flavor to the discussion.

The study is called the UK Employment Retention and Advancement Demonstration, or UK ERA, which was influenced by the ERA study that MDRC conducted for HHS. In the UK, MDRC led a research consortium to design and evaluate the UK version.

The program intervention was targeted to three types of people: unemployed lone parents, similar to TANF recipients; lone parents working part-time and receiving the UK equivalent of the Earned Income Tax Credit (EITC); and a group of long-term unemployed people, mostly men, a group that receives little cash assistance in the US.

The study was designed to test the effects of adding a two-year post-employment component, which consisted of a combination of in-work job coaching and financial incentives for full-time work and skills training, to Britain's regular welfare-to-work program, which was primarily a work-first job-placement program. The question was whether the extra post-placement support and

incentives help participants work more steadily and advance to better jobs. The intervention, from the time of random assignment to the end of the post-placement support, lasted 33 months.

Because it was important to determine whether the program had positive effects on employment and earnings and reduced welfare receipt after the intervention as well as during it, program and control group members were followed for five years after random assignment, or more than two years after the intervention ended. This timeframe was important because advancement in work doesn't happen quickly.

The evaluation was comprehensive. Over 16,000 people were randomly assigned across nearly 60 welfare-to-work offices in six different regions of the country. And it was an expensive evaluation – in the millions of pounds – due to a number of factors.

- The large sample and number of locations.
- The intake interview took longer, because of the extra time need to explain the new study, get informed consent, and conduct random assignment.
- The random assignment module and procedures needed to be built to make sure the allocation of people was truly random and could not be gamed by the intake workers – in other words, they could not affect who would end up in the program or control group.
- Intake workers were trained on random assignment, and research staff monitored the process for the year during which it occurred.
- Sites were offered program-related technical assistance, to help them understand the model and increase the chances that they would implement it properly. This was important because the evaluators wanted to make sure they ended up giving the "concept" of UK ERA a fair test.
- The implementation study required multiple site visits in order to learn whether the model on paper was actually implemented in practice – whether they were actually testing in the evaluation what they thought they were testing, and how implementation of the program may have varied across target groups and places.
- The data collection effort was substantial, including the collection of administrative records data for employment, earnings, and welfare and other benefit receipt, and three waves of surveys for subsamples of program and control group members.

Surveys are expensive, but are also necessary for these types of evaluations. In addition to determining whether the program increased work and earnings and reduced welfare receipt, it was important to determine whether participants got better jobs over time and whether more income had spillover effects into other areas of their lives – for example, reduced material hardship, or had positive effects for their children. It was also important to follow people over a long enough period of time to determine whether there were effects while the program was operating and the extra support and incentives were in place, as well as afterward, after those supports and incentives ended. Survey data also provided information on the services and supports members of the control group accessed on their own from other sources in the community. This information was needed to understand the nature and size of the intervention we were actually testing and the treatment/control differential.

And, of course, the research team needed to process all the data they collected, ensure that the data protections were maintained, analyze the results, and write reports.

Could the research team have done this evaluation less expensively? Yes. But, of course, there are tradeoffs.

- They could have used a much smaller sample. For example, they could have tested the program for only one group, not three. Or they could have had smaller samples for each group, or not used so many locations. If they did any of these, they would have less evidence about the program's effectiveness across subgroups of people and places, and depending on the ultimate sample size, less ability to detect with confidence smaller but still policy-relevant effects.
- They could have collected less data. For example, they could have limited the study to just administrative records on earnings and/or welfare payments, and not administered the surveys. But that would have left important questions about control group services and potential effects on advancement and quality of life unanswered, and precluded a reliable cost-benefit analysis. It would have precluded learning about participants' training experiences and skills acquisition.
- They could have followed participants for a shorter amount of time. But then they wouldn't know whether the program's effects were sustained, dissipated, or even grew after the treatment ended.
- They could have skipped the implementation research. But then they wouldn't be sure what they were actually testing.
- They could have skipped the cost-benefit analysis. But that would have left an important policy question unanswered.
- They could have skipped the programmatic technical assistance. But that would have risked not testing what they set out to test.

Fortunately, they did not have to make these choices as this was a very well-funded study. But it was launched it at a time when the British government had money. Today, a study of this magnitude would be almost impossible in the UK, given the current fiscal climate.

So how do we continue to build evidence of important innovations?

With UK ERA, they could have conducted a much smaller and much less comprehensive evaluation and still learned valuable lessons. One alternative would have been to scale back and not try to learn everything in one study. For example, as mentioned earlier, they could have focused on a smaller sample (e.g., a sample of 6,000 rather than 16,000), either by conducting the study in fewer locations or with fewer target groups. They could have eliminated the surveys and cost-benefit analyses, and reduced the scope of the implementation research. They could have tried to answer fewer questions, focusing on the bottom-line inquiry, such as "did the program improve employment and earnings and reduce welfare receipt?" If they found no positive effect on this "bottom-line" outcome, that would suggest that the approach was probably not worth further investment in the same form. On the other hand, if larger effects were observed, that might justify

continuing to invest in evaluating this intervention – perhaps with a series of small, future replication studies to answer other important questions. The learning agenda could be staged starting with small-scale RCTs and moving toward larger-scale studies if the findings warrant.

So researchers and policy makers shouldn't be scared off from RCTs. Choices would have to be made on the scope of the learning agenda, but it would be incorrect to conclude that random assignment is the crux of the problem. Indeed, some of these same choices would have to be made even for non-RCTs, if they involved comparison groups with comprehensive data collection.

Getting the Policy Answers You Need: The Latest in How Randomized Evaluations Can Inform State/Local Policy Decisions

Steve Bell then commented on work underway in the Department of Education and reflected on methodological issues.

Bell noted that the Department of Education has supported over 70 local evaluations of education innovations – known as Investing in Innovation studies, or "I3." Grantees are conducting evaluations to contribute to evidence of promising practices, effective programs, and scale-up methods. Abt Associates supports study rigor by working with grantees on the evaluation designs. About half of these evaluations include RCTs. Similar approaches (a national evaluator working with individual site evaluators) are underway at the Department of Labor and the Corporation for National and Community Service.

Bell went on to note that RCTs can go beyond providing a verdict on whether a program has impacts to helping the field understand what about a program works. Using advanced analytic techniques and design can help us get inside the "black box" to assess the effectiveness of specific policy and program components. Bell also described strategies for making RCTs "easier to take" by engaging local staff in design decisions and looking for ways to decrease local burden associated with participating in an RCT. Finally, he noted that RCTs can be implemented "cheaper, faster, better – no kidding" by using both local and national administrative data.

Random Thoughts on Random Assignment: Six Lessons Learned

As a former state administrator, Don Winstead participated in two large RCTs in Florida in the 1990s. He shared his insights in this final presentation of the plenary session.

- **Project Independence** – The evaluation included nine Florida counties randomly selected from 25 candidate counties. Within these counties, 18,239 single parents were randomly assigned to a treatment group who got the Project Independence welfare-to-work program and a control group who got standard services that would have been available in the absence of the program.
- **Family Transition Program** (FTP) – The evaluation included 2,800 participants in one of the first time-limited welfare-to-work demonstrations. Half of the participants were randomly assigned to the FTP group, which was subject to time limits, rigorous work requirements, and enhanced case management and supportive services compared to a control group that received the pre-FTP services.

Winstead, therefore, approaches the use of RCTs from the perspective of the "lab rat" in these experiments and provided the following lessons learned:

1. Rigorous Evaluation is Not Horseshoes. You don't get points for being close to rigorous. There is a tendency to think of evaluations methodologies on a continuum. RCT experiments are at one end – the gold standard. This is followed by quasi-experimental approaches like propensity matching or a variety of other methods such as county-to-county comparisons, other analytic approaches like multivariate regression using administrative data, outcome-based approaches, and, at the other end of the continuum, anecdote and hunch (the aluminum standard?). Winstead notes that quasi-experimental methods do not necessarily approximate RCTs. They can be an attractive alternative, but not as reliable as experimental methods (though non-experimental approaches can still have significant value).

2. All That's Gold Does Not Glitter – or put another way – the gold standard is not a silver bullet.
    - Even the most rigorous methodology can still have shortcomings. Frequently, demonstrations include multiple service interventions or multiple policy changes. This creates a "black box" and it's hard to know which of the changes was responsible for any observed impacts.
    - It can also be difficult to avoid "contamination," particularly in highly visible reforms. For example, in the FTP, there was an incredible amount of publicity – nationally, internationally, and, of course, in the local community. There were visits by the BBC, Japanese TV, and several national news outlets including the *New York Times* and the TV networks. When the first person reached the time limit, it was a front-page story, above the fold, in the local newspaper. Consequently, some number of people in the control group thought they were subject to time limits. This can affect the measured impacts of the program.

3. Don't Neglect Implementation Studies (also called process evaluations). RCTs can indicate whether a program had impacts but, particularly if it includes multiple policy changes or service components, it is difficult to know **how** and **why** there was or wasn't an impact. A well done implementation study can clarify why the program worked or didn't. As a program administrator, Winstead had a tendency to think that when he signed a memo or policy directive, change inevitably happened all around the state. Sometimes the policy or program may not produce the desired result because it was simply not implemented as expected. A well done implementation study is a critical part of any evaluation of significant size or scope.

4. Feedback is the True Breakfast of Champions. Sometimes it can take years to get the results of RCTs. The cycle of rigorous research can be longer than the political/programmatic cycle of change, so don't wait for the end of the study to get findings. It is important for operational purposes to get regular feedback. In the FTP, Winstead developed a system in

which the state agency gathered identifiers on the people randomly assigned and kept a running tally of the cash assistance and food stamp benefits as well as the earnings of the program participants (without ever sharing this information with the local sites running the program). That way the state could monitor the progress of the demonstration in near-real time and not wait five years to learn if it was successful. When the state had to make critical decisions about implementing state policies when national welfare reform passed, the agency had the information to determine, for example, that it was cost-effective to implement an enhanced earnings disregard. The state had detailed information on the impact of this decision although the final evaluation report was not due for another two years.

5. There Really is No Substitute for Experience and Expertise. Designing and implementing complex RCTs is not for beginners. Winstead's advice is to design your procurement so that you get an evaluator that has done it in the past in the real world.

6. Know Why You Are Doing the Research – But Use the Available Tools. Historically, there are two reasons to conduct rigorous RCT evaluations:
    - To learn what works and how to improve programs and services, and
    - To get a federal waiver.

    In 1996, the second reason went away for TANF, but lately it has reemerged. Policy makers may want to do an evaluation to learn what works, but may not have funding or support. The federal waiver requirement can be an important marketing tool in furthering the goal of learning what works.

Examples of Small-Scale Randomized Control Trials

In the first Research Academy breakout session, facilitated by Michael Wiseman from George Washington University, participants learned about examples of small-scale RTCs in both private and public settings. Speakers included Richard Hendra of MDRC, Kinsey Dinan from the Office of Evaluation and Research in the New York City Human Resources Administration (HRA), and Erik Beecroft from the Virginia Department of Social Service.

What Can We Learn From Experimentation in Business?

Richard Hendra built on Jim Riccio's plenary discussion points by relating his personal experience in the market research field. He noted that businesses have developed the practice of conducting thousands of large-scale, short-term RCTs to identify the most effective methods for reaching and engaging clients. This continuous evaluation enables faster evolution of practices as companies integrate findings into their marketing strategies.

To illustrate the types of studies conducted by marketing departments, Hendra walked through a study that was designed for a grocery store. Data on shoppers is regularly collected through loyalty cards. This dynamic dataset is used to advertise alternatives. Three groups were created to disentangle the messaging preferences of different types of shoppers. These targeted messages separated typical "demographic" shoppers, organic shoppers, and "heart healthy" shoppers based

on their purchase history. Each target was then split into two treatment groups and a control group, for a total of nine research groups of 35,000 shoppers. One treatment group received a current advertisement, the second treatment group received an alternative ad chosen by previous copy testing, and the control group received no advertisement. Over the course of four to eight weeks, shoppers were tracked for their purchase of the advertised product. Researchers could then compare the results from each research group to determine which marketing method was most effective for each of the target groups.

Hendra discussed challenges that social policy experimentation faces that are less relevant in a business marketing setting:

- Less control over the environment;
- Higher cost of data (surveys, accessing administrative data, etc.);
- Longer period necessary to allow outcomes to emerge;
- Longer periods necessary to recruit sample; and
- Effort needed to convince individuals to participate in random assignment.

However, he noted that the methods employed for conducting RCTs in the private sector were virtually identical to those used in social policy experimentation.

Hendra also noted that there are many potential pitfalls of random assignment, such as attrition and poor data preparation. He stressed that when a policy organization starts to think about random assignment, staff should seek advice from experienced researchers and research firms to ensure that the process is ultimately beneficial.

Despite these unique challenges, Hendra shared his belief that policy organizations could move closer to the business world model of continuous, integrated evaluation. Necessary conditions include the flexibility to innovate, the development of research questions with a narrow focus, and the collection of data that can contribute to RCTs.

CASE EXAMPLE: New York City Child Support Program Pilot

Kinsey Dinan then shared an example of how she integrated random assignment into the implementation of a pilot program for New York City (NYC) Child Support.

The Cash Assistance Agreement Pilot (CAAP) was designed to address the low order establishment rate through Family Court, thought to be the result of many non-custodial parents' aversion to the court process. CAAP offers an alternative order establishment process that bypasses the court by developing an order agreement between the custodial and non-custodial parents under supervision of Child Support Customer Service staff. If an agreement is signed, parents don't have to go to court.

Dinan outlined three key steps that had to occur before the RCT was implemented.

1. Define Key Pilot Outcomes. In the CAAP study, the goals were to increase the number of child support orders, improve compliance with orders, and reduce the number of days between court referral and order establishment.

2. Identify Appropriate Evaluation Design. The Office of Evaluation and Research determined that an RCT was the most reliable and easiest to implement. While pre/post-test and comparison group designs were briefly considered, Dinan and her group decided that ongoing changes within the agency made it too difficult to make reliable comparisons across time or location.
3. Make Your Case to Program Staff. Dinan stressed the importance of gaining buy-in from program staff and comprehensively explaining the purpose and process of random assignment. Program leadership also need to commit to the evaluation and express their support. In this study, Dinan found consensus with program staff by addressing their concerns of how many participants were assigned to the control group and when random assignment would begin in the new pilot program. The office decided on a 60/40 split between treatment and control group members, and random assignment didn't start until the third month of pilot implementation.

The Office of Evaluation and Research randomly assigned non-custodial parents to either go through the standard Family Court order establishment process or participate in CAAP and found that the random assignment process itself was a small burden compared to the time and resources required to implement the pilot program itself. The office used data already collected by the agency to assess the impacts of CAAP and found that there was not a difference in the order rate for the treatment and control groups. However, interviews with staff and non-custodial parents suggest that the pilot experience was perceived as positive. In 2012, CAAP was expanded to all NYC boroughs.

Dinan concluded by describing possibilities for future research in the CAAP program, including analyzing the other two key pilot outcomes, conducting a cost analysis, and developing a process evaluation of the program.

CASE EXAMPLE: Earned Income Tax Credit Take-up

In the final presentation of this session, Erik Beecroft discussed random assignment strategies used to study the take-up rate of the EITC among clients of the Virginia Department of Social Services (VDSS). State law requires VDSS to notify clients with earnings about the federal EITC. In 2006, a study by the Virginia Joint Legislative and Audit Review Commission estimated EITC participation for this population was only 13%. As a result, state law was changed to enable the state Department of Taxation to share data on federal EITC receipt with VDSS. VDSS began conducting semi-annual data matches to identify eligible non-claimants. VDSS also initiated two outreach strategies (postcards and automated phone calls) and Beecroft proposed using random assignment to determine whether this outreach was effective. The EITC outreach coordinator, when approached, was not interested in using random assignment because the control group would not receive the innovation. However, Beecroft persuaded management on the value of knowing whether outreach is effective, and got permission to do the experiment.

Beecroft used administrative data from state individual tax forms for this study. He targeted families who receive public assistance and have earnings, but did not claim the EITC in the prior tax year. In the first year, he looked at households that received either a mailing, an automated phone

call and a mailing, or neither form of outreach. In the second year of the study, he further divided the population into those receiving only a mailing, only a phone call, both a mailing and a phone call, or nothing. In both cases, he found small positive impacts on filing state tax returns and claiming federal EITC dollars when individuals were both sent a mailing and received an automated phone call. Beecroft also conducted a cost-benefit analysis to determine if these means of outreach are cost effective.

Beecroft used this example of EITC claimants to illustrate that under the right conditions, random assignment can be an easy and inexpensive way to determine effectiveness of a policy or strategy. He shared other possible circumstances for using random assignment, including changing wording on surveys to affect response rates, IT changes such as automating income verification on TANF applications, and conducting outreach campaigns such as using Supplemental Nutrition Assistance Program (SNAP) dollars at farmers' markets. In addition, he touched on some challenges that can arise in this type of study, including:

- Senior leadership not understanding the difference in credibility between random assignment and other study designs;
- Demands on staff time; and
- Risks to program managers if an experiment does not provide evidence that a program is effective: making significant program changes, losing funds, and staff.

Questions from the Session

1. Do you have perspectives on overcoming confidentiality issues such as informed consent in your research?
    a. Hendra noted that businesses do not have this concern as much as social programs, and that companies are careful to avoid personally identifiable information. In the marketing industry, loyalty cards are preferred to collect data independent of personal information.
    b. Dinan noted that privacy is already a concern in social service agencies and that there is no difference when running a study. The goal is to use data that are already collected for an additional purpose, while keeping privacy and data security in mind. When you contract out for portions of the study, it may be a challenge to share this data easily and securely. She noted that informed consent should always be asked of participants.

2. Given the results of the study at the NYC Child Support Office, it is surprising that the pilot was expanded. Couldn't you have looked at other outcomes before expanding the program? I'm hesitant about random assignment because some legislators will do what they want even if the results suggest something different.
    a. Dinan explained that there were several goals for the pilot program, one of which was to change the face of child support in the city and make it friendlier. The Office of Evaluation and Research believes that, over time, this program could help to do that.

b. No impact in random assignment is often enormously informative. To know that the program is doing no harm and is moving in a direction that staff appreciate is a useful outcome.

3. Staff in NYC felt that random assignment was a relatively small lift. In our county it's seen as difficult. What was your approach to training and communicating with staff?
    a. Dinan believes that the 60/40 split between treatment and control helped to explain to staff why they couldn't assign every other participant into the treatment group. The office used a spreadsheet to conduct random assignment and integrated this file into training. They also used two colors of folders to physically separate the cases. Staff felt that implementing random assignment was simple compared to the task of executing the pilot program itself.

4. Controlled experiments seem difficult to implement considering all of the small pieces in a program that must be coordinated across a county or counties.
    a. Some changes are not difficult to implement – such as a change to an IT system or outreach intervention such as a letter. Policy changes in which clients are treated differently become harder to do.
    b. Dinan emphasized that in her study, the existing process as well as the new pilot process were both going to be in place, in part because some cases had to use the traditional system and in part because the office was planning to roll out the pilot on a staged basis. In this instance, there was a window of opportunity to make random assignment happen because clients were already being sent to one of two treatments.

Using Small-Scale Randomized Control Trials

The second breakout session was designed as a workshop to engage participants and think through the necessary components of small-scale RCTs. The session was led by LaDonna Pavetti from the Center for Budget and Policy Priorities, who started by asking the group, "What do you think the characteristics are of good ideas for small-scale random assignment evaluations?" Answers included the following:

- Readily available or easy-to-obtain data that will help answer the question you are asking;
- Ease of assigning clients to treatment or control groups;
- Large population of clients receiving or requiring services;
- Procedures to keep programs or policies distinct between treatment and control groups;
- Study sample representative of the population;
- Test of something easy to manipulate;
- Test of something already happening, even if it is difficult to implement (for example, testing impacts of a new pilot program);
- Ability to develop differential treatments free of ethical issues about withholding services;
- Willingness of the staff and organization to change their practices; and

- Test of something meaningful that still meets the above criteria.

Pavetti then shared an example of a random assignment study that met these characteristics. In this example, a company tested the inclusion of a rating on customers' electric bills that show where they fall in energy use in relation to their neighbors and the general population. The company theorized that this rating might reduce energy use. In the study, the control group received an electric bill as it normally looked, and the treatment group received a bill that included the rating. After analyzing data on the energy use of these two groups, the company found that the rating did significantly reduce electricity consumption among the treatment group.

One participant in the session then shared his idea for a random assignment evaluation in his Department of Child Support (DCS) office. He suggested using different letters to persuade non-custodial parents to contact DCS with the goal of reducing sanctions or increasing the sanction reengagement rate. He would use performance measures that are already collected to look at the sanction reengagement rate (those non-custodial parents who were sanctioned in the previous month and engaged in the present month). This participant would randomly assign non-custodial parents to receive different letters.

This example generated a discussion around what should be measured. The initial suggestion, the sanction reengagement rate, is easy to measure but is further removed from the letter. In this case, the hope is that a non-custodial parent would call a case manager based on the letter and ultimately reengage. The suggested alternative was to directly measure the number of phone calls to DCS generated from the letters. However, this approach is more challenging to measure, requiring either a change in staff practice or a process of logging calls on a dedicated phone line. The group concluded that the initial measurement is appropriate and that the office can further tease apart points in this flow if the first test is inconclusive.

Other examples of potential random assignment studies included the following:

- Testing the impact of online benefit application designs on the completion rates of applications, where participants are randomly assigned to alternative design versions when they click a link to the application.
- Varying the type of communication between child welfare and TANF case workers for children who transition from foster care to permanent independent living. One state volunteered that they use varying practices across the state and could possibly test these different communication strategies using randomly assigned pairs within one or more counties.
- Randomizing which contracted providers clients are sent to, and possibly digitizing this process, to evaluate the differences in contractors. For example, NYC job center clients will be randomly assigned to one vendor of a geographically similar pair. Pavetti suggested structuring the study to have a narrowed research question while others expressed that even if there is not a specific research question, differences that arise among contractors could lead a department to further examine a program.

- Testing how well a vendor's soft skills training program works. Pavetti suggested one possible study design where the control group is in job search activities while the treatment group received social skills training before going into a job search. The two groups' success in finding employment could provide some data on whether the soft skills training was effective.
- Measuring the impacts of long assessments to see if more basic assessments are equally effective. Participants were reminded that it is also possible to remove a component of a program or service and test the impact using random assignment, just as they can add or change a piece.

During this exchange, several other points were discussed. Pavetti emphasized that if an organization wants to study something big or complex, it is best to hire a firm that specializes in random assignment, giving the example of a federal waiver experiment. However, she noted that a lot can be done with the data that agencies collect, which is easier than ever to work with because it is largely automated. Pavetti also acknowledged that there is middle ground between these large-scale studies and in-house data analysis. Organizations can hire a firm to provide support for collecting informed consent, addressing data issues, and ensuring the ethical treatment of human subjects.

Using Randomized Control Trials to Learn What Works

In the final breakout session of the Research Academy, facilitated by Brendan Kelly from ACF, presenters Mike Fishman of MEF Associates and Karen Gardiner of Abt Associates reviewed the characteristics of both small-scale and large-scale random assignment studies. These presentations were followed by an interactive session focused on the question, "What resources would be helpful to you in feeling comfortable and willing to take on something like this?"

Using Small-Scale Random Assignment Experiments to Learn What Works

Fishman reviewed the concepts and themes of the Research Academy in his presentation. He began by reminding the participants of the difference between outcomes and impacts and stressing the "gold standard" of random assignment evaluations for illustrating the true impact of programs and strategies. He countered the common belief that random assignment studies are expensive and complicated, drawing on points mentioned throughout the Research Academy. Fishman noted that administrative data can be used for the analyses, that the impact analysis doesn't require complex statistics, and that the random assignment process can be automated using Microsoft Excel or an online random number generator. He also noted that this type of study does not have to last a long time; the duration is dependent on the time required to build a sample, how long it takes to implement the intervention, and the follow-up period needed to assess the impacts. However, he emphasized that RCTs have to be designed with care to ensure that the proper impacts are being properly measured and can be understood.

Next, Fishman walked the participants through the key steps to successfully executing a random assignment study. He noted that the way these steps are implemented depends on the scale of what

is being studied. However, these should be looked at as things to think about during the study process.

- Define research question(s) that will inform each subsequent step of the process.
- Define treatment and control group services. The control group will receive alternative services, generally conceptualized as what an individual would receive in the absence of the treatment. In studies of social services, the control group is rarely equivalent to no service. The control group could also be a group that is getting the current service, and the treatment group is receiving an enhancement or change to that service.
- Determine sample size needed to assess impacts. In addition to analyses on the entire sample to determine the overall impacts of the program, subgroup analysis can provide specific information about the impacts of the program on particular types of participants. However, with a smaller subgroup you can only see larger effects, so the question of power and sample size remains.
- Develop a client flow model to understand how the program works now and how it will change with the study. Questions to answer when developing the client flow model include:
    - Where do participants come from?
    - How do they currently move through the service delivery process?
    - Where should random assignment take place to best capture the effects of the program being tested?
    - What are post-random assignment flows for treatment and control group members? Everything that happens after random assignment can influence the impacts that you see in the study. Once people are in the program group, they are always in the program group regardless of how much or little they engaged with the program. The same is true for the control group.
- Identify data to be collected at baseline and for outcomes using existing administrative data wherever possible. In order to be useful, these baseline data must be current to when the client enrolls in the program.
- Develop informed consent and obtain Institutional Review Board (IRB) review, if needed. If the organization does not have a lot of experience in research, it may be helpful to work with an outside organization to ensure that this process is completed thoroughly.
- Develop random assignment procedures manual and train staff in random assignment procedures as appropriate. This training should have precision and clarity to communicate the importance of random assignment and enable staff to accurately communicate about the study to potential participants. Random assignment training should be as close to the start of random assignment as possible.
- Begin random assignment and monitor the process to ensure it is being implemented with integrity.
- Conduct an early assessment of the study, including random assignment procedures, treatment group services, and control group services to ensure that random assignment is being implemented properly, that program services are being delivered as planned, and to understand the nature of services being received by control group members.

- Conduct implementation research to understand the steps the clients take along the way in the treatment group. The control group is also important to understand in the implementation research – are they finding similar services to the treatment group somewhere else? This could reduce the impacts that you see in the results.
- Conduct the impact analysis. For RCTs, this may be as simple as a comparison of means for program and control groups. While there may be some additional statistical corrections, experimental impact analyses are generally simpler than more complex quasi-experimental designs.
- Finally, a cost-benefit study can enable an assessment of how the benefits derived from the program compare to the costs. Once there are impact results for key outcomes, such a study may be relatively simple to implement.

Example from an ISIS Site

Following these comments, Karen Gardiner, Project Director for the ISIS study at Abt Associates, discussed how this study was designed and is being implemented at a program (Valley Initiative for Development and Advancement, or VIDA) in the Lower Rio Grande Valley in Texas. She noted that ISIS is a large study, but the steps its team goes through at each site are similar to what happens in a smaller study.

The VIDA program works to train participants for local jobs. Most participants are low-income, Latino, and single parents. They may be unemployed, underemployed, or on public assistance. The program uses a "Career Pathways" model to provide tuition, other financial assistance, and counseling to help its clients progress in their career of choice and reach self-sufficiency. VIDA leadership welcomed the ISIS study because they want to demonstrate program impact on individuals in an official and rigorous manner. This, they hope, will show to funders that the program "works" and thus will open up an opportunity to scale up and serve more participants. Gardiner noted that the national attention brought by the ISIS study was seen positively by the program.

Research questions used for the VIDA program had already been developed by the ISIS project: What is the impact of the program on persistence in education and achievement of credentials and degrees? What is the impact of the program on career-track employment and earnings? What is the impact on well-being? Although the questions were in place, the ISIS team worked closely with the VIDA team to understand what they meant for the program. Gardiner emphasized that it is important to make sites feel that evaluation isn't something done to them, but something in which they are a willing participant.

Over a two-year time period, the program will randomly assign at least 1,000 individuals to treatment and control groups. The control group cannot be a part of the program but its members are able to access other services in the community. Gardiner explained that ISIS will produce site-specific results from the impact and implementation studies, and will also feature a site-specific cost-benefit study.

Gardiner then walked through the steps the study team and program took to prepare for the study and for random assignment.

1) Engage stakeholders, including local Workforce Investment Boards, local and state politicians, and representatives from local colleges and partner organizations. It is important that all of these stakeholders understand why the study is important and feel that they have a role to play. To facilitate this, the site hosted a kick-off event before ISIS started in order to spread enthusiasm about the study and to get the attention of the community.
2) Rework outreach and refocus messaging to reflect the study. The VIDA program had to brainstorm new potential referral sources, as it had to recruit twice as many participants as would be enrolled in the program. It chose to market the program via print, TV, radio, and word-of-mouth and tracked the number of interested individuals coming in from each tool to learn where best to invest the program's time and resources. In addition, the ISIS study and random assignment process had to be incorporated into the messaging about the program. It is important from an ethical perspective that everyone knows as early as possible that the program is involved in the study and admitting participants based on a lottery-like system.
3) Incorporate evaluation data collection into the eligibility process. The ISIS study has several data collection forms that were to be administered prior to random assignment. The study team worked with the program to determine the best method to collect this information. More generally, it was critical that the study and program staff had the same understanding of the participant flow through the intake, random assignment, and enrollment processes. The ISIS team developed a flow chart to depict this process in a way that was easy to understand.
4) Staff complete random assignment training and conduct random assignment with eligible individuals. Random assignment is ongoing. Based on their random assignment status, treatment group members begin to meet with their counselors to begin the program; control group members are given a referral list for other programs in the community. Control group members are barred from enrolling in the program for two years.

<u>Comments and Questions from the Session</u>

What kinds of resources would be helpful to you in feeling comfortable and willing to take something like this on?

- "The Research Academy is a great support. It is really bringing administrators into the world of research. I hope it continues because it makes [research] seem more accessible to administrators of programs. I hope there are some post-NAWRS meetings to keep the discussion going"
- "The Research Academy has been a great opportunity to listen to everyone. We have been trying to think through different service delivery options in light of whether a waiver happens or not and this has made me think through how to conduct random assignment within a diverse, county-based service delivery system. How would you approach conducting these kinds of tests in a county-based system?"

    - Fishman suggested that, in this setting, it would be helpful to engage with the counties to learn whether they are looking at alternative ways to operate and what they would do differently. Learning what they want to test and improve would ensure the study was relevant and also gain buy-in. Larger counties may be able to do an independent test, while smaller groups of counties may be doing similar things and could work together to test something and pool the results. You have to start with county feedback and see what comes out of the discussion.
- "I'm thinking about the back end and what happens when you get the results of the study. How do you explain and talk about the results when you don't get findings?"
    - Policy makers and practitioners learn from experiments whether or not they show positive impacts. If you don't get positive impacts from one test, you should try something else. Small-scale evaluations are good in this regard because they can provide more immediate feedback. The goal is to improve program effectiveness. RCTs could become a normal part of doing business, just as performance measurement has become. This is what is happening in the private sector. When something doesn't work you can learn about why it doesn't work and use that information to try something else. If the results are positive, it is important to make sure that the people operating the program know about it and continue to do what they are doing.
    - Brendan Kelly noted that research is often framed as a test of an additive property, when you can also think of it as finding ways to produce the same outcomes by doing less, so that you have more money available to put toward other things. Conducting a study that shows no impacts might say that what you were initially doing was adequate and you should focus attention somewhere else.
- "It is helpful to have a network and resource for those who aren't as familiar with research…where people could share concerns about their study design and examples of what they were doing. I don't know many other local and state evaluators around the country [who I could reach out to]."
- "Some things are very different in different states, but I'd like to see us learning from each other on these small-scale programs to add to each other's' efforts instead of replicating the development process. People could share the results of their evaluations and you can consider if it may or may not be different in your location. We could develop stronger models based on other experiences."