Lessons Learned from SSA Demonstrations: A State of the Science Meeting

June 15, 2021

Transcript of Panel A: Design and Use of Demonstrations
Panel Chair: Laura Peck, Abt Associates

10:10 – 10:45 a.m. EDT: Design of Demonstrations
Presenters: Burt Barnow, The George Washington University
David Greenberg, University of Maryland, Baltimore County
Discussants: Jesse Rothstein, University of California, Berkeley
Jack Smalligan, The Urban Institute

Thank you, Austin and of course, one minute ago my Internet just went out. So for now, you have only my voice and not my face, but know that I'm smiling at all of you and welcoming you to this really special event. I think that if there is a silver lining to our pandemic state, it is that this meeting can reach so many more people than only those who would be able to show up in DC and attend in person. And I even saw on the invitation list, someone from A New Lease, which is a local nonprofit, where I live in Phoenix, Arizona that I volunteer for so welcome to them and welcome to all of you and from all corners of the country. This 1st panel has 2 presentations each with a discussion, and the first session is on the design of the evaluations that are part of demonstrations, Burt Barnow and David Greenberg crafted an insightful paper on these design issues and they will present their key findings over the next 15 minutes followed by discussion from Jessie, Rothstein and Jack Smalligan. As Austin explained, after the discussions, we will have a brief break and then a second presentation with discussions after which we'll get to those questions that you can drop into the Q&A box, so now I will turn it over to Burt Barnow.

Thanks so much for being here. Thank you, Laura. My name is Burt Barnow from George Washington University, and I'm going to do the presentation of the design of Social Security Administration demonstration evaluations. I would like to start by thanking the Social Security Administration and Abt Associates for inviting us to this most important and interesting conference. This paper is co-authored with David Greenberg and it's about how the demonstration evaluations were designed at the highest level. What we aim to do is consider how best to evaluate interventions and to develop changes to policy and program. So we know what their impacts are and whether they should be rolled out at the national level. We reviewed 16 evaluations, which this presentation is based on. Next slide, please. This slide shows the types of evaluations that we have, you can see that we have non-experimental and experimental evaluations. I'll get into that in a second. And on the right side of the slide, we have examples which I won't be going through for time purposes. The SSA evaluations involve these design types, and the two non-experimental types are called proof of concept studies and non-experimental impact evaluation. A proof of concept study, simply assesses if an intervention can be successfully implemented with little emphasis on estimating its impact non experimental impact evaluations. Assess the impacts of the program, and they have to rely on assumptions to control for differences between the treatment group and what are called comparison groups that do not get the treatment.

We encourage the Social Security Administration to continue conducting proof of concept studies. They should be clearly labeled as such so that people aren't looking for impacts when they look at them. And

if successful, of course, more rigorous evaluations can be conducted. We believe that non experimental impact evaluations that were implemented among the 16 had some limitations, probably because the decision to conduct the evaluations was not made until after the programs were implemented. Most of the designs were experimental in nature where assignment to the treatment, or control group was done with random assignment. And when that's done, the impact of the program can be assessed, simply as the difference, and the outcomes between the 2 groups. There was also 1 that is what's called a natural experiment where although assignment wasn't random, it was close to it because it was based on the last digit of people's social security numbers, which is close to being random. Next slide, please. Population representativeness among major national evaluations SSA's demonstrations are strong and their efforts to ensure population representatives this is important for the results of the evaluations to apply to the overall population of interest and that's referred to as external validity.

Some ways that, that given evaluation, or given demonstration may not have external validity, is that the, the experiment, or the demonstration may be carried out in limited sites. And they may not be representative of the entire nation, because the characteristics of the recipients or economic conditions may not be representative. Another problem is that. Demonstrations using volunteers may not be representative of what would happen if the program became mandatory and as we heard in Jeffrey's introduction has limitations on doing things that are mandatory. Finally, demonstrations that focus on specific medical conditions like diabetes or mental health problems may not generalize to findings for other conditions so these observations lead to the conclusion that evaluation should always consider not only whether the evaluation has what we call internal validity that it gives an accurate assessment of what happened, but also external validity and applies to the entire population of interest. Next slide, please. Statistical power and effect sizes. So this is a little bit wonky. I'm sorry to say, I just discussed broad issues related to evaluation designs and their ability to support causal claims and to be generalized. This part of the evaluation focuses on so narrower technical decisions about executing evaluations and practice. I will talk first about statistical power analysis. And then outcome measures and impact estimation issues. The first of these technical topics is an evaluation's statistical power to detect impact. Most of the evaluations indicated that they conducted a power analysis in designing the evaluation. But the details, we're not always provided in the reports. Having too small of a sample means that the evaluation might not be able to detect effects of a policy relevant magnitude with a high enough level of statistical significance so that we can be confident of the findings.

Some evaluation cited costs is the reason for having a smaller sample, as the cost sometimes turned out to be much higher than was anticipated. Even with a sufficient overall sample size, we might be interested in the impacts on specific subgroups, and so evaluation should be planned to achieve sufficient sample sizes for priority sub groups. On the slide, I give an example of one that had that. So, our conclusion here is that the power analysis is a very important thing to conduct in designing the evaluations so that the results are credible. Next slide, please. Outcome measures. The most common outcome measures that were used in just about all the studies were employment and earnings usually whether or not people work. Sometimes hours of work and earnings was generally measured annually the studies mostly measured, total earnings. But sometimes they measured whether earnings exceeded substantial gainful employment, which is used in determining continuing eligibility.

Another outcome that was sometimes measured was benefit receipt and amount and some evaluations use time for processing applications. We thought a limitation of these evaluations is that they did not

always include the accuracy of the determinations. Some evaluations used health outcome measures, including mortality and doing so they use self-assessment or quick measures, such as the Mini, mental state evaluation, and the mental health inventory. Some evaluations use surveys to capture other outcomes that aren't available in administrative data. That includes hours worked, other unearned income, hourly wage rates, motivation, quality of life, health insurance, receipt of specific services, understanding the rules of the demonstration. So, conclusions here, evaluations, focused on the most important outcomes, and we thought it might be useful for SSA to standardize the measures of health specifically where there are a lot of potential measures available.

Next slide, please. So impact evaluation issues, the. SSA studies, almost all, relied on what's called an intent to treat impact measure that simply means that they compared the outcomes for people who were assigned to receive the treatment to the outcomes for people in the control or comparison group. I'll get back to it that in a minute with one of the alternatives. The methods used were typically ordinary least squares for continuous outcomes and logistic analysis or logit for dichotomous outcomes. The studies used weights sometimes without explanation or sensitivity analysis in their analysis in their evaluations, and they use data from both surveys and administrative sources. So, some comments on these issues. We suggest that consideration be given the estimating the impact of the treatment on the treated sometimes called TOT which requires more assumptions, then the intent to treat approach, but often answers important policy issues. With respect to weighting strategies, we think more work needs to be done in this topic and it is somewhat controversial. We suggest that the evaluations conduct sensitivity analysis on their weighting strategies and explain what their preferred strategy is and why. There is substantial literature on how surveys and administrative data can lead to different findings and given that these evaluations commonly use both, there's the potential for first, understanding better the interventions impact by seeing how the impacts differ across databases. But also, for informing the literature about the conditions, under which survey or administrative data provide better estimates. Next slide. Several additional data issues have implications for the findings.

So, for example, the follow up periods varied among the studies. Some of the studies looked at very short term events like, for example, improving the process for application. And there the follow up was relatively brief, a year or less, but most of the evaluations looked at the issue of whether or not these particular demonstrations were able to increase employment and earnings over the long run. And in those cases, the fellow periods lasted 5 years, 10 years are often longer. That's not difficult to do because you can use the social security earnings data, which is automatically collected for all people in the country. There were 2 types of missing data that were accounted for, in the studies when an entire unit didn't respond. If someone was missing that was dealt with, by weighting the observations to take account of the prevalence of different types of people. And if individual items were missing, that was usually dealt with by imputing the missing data. Two of the evaluation dealt explicitly with what is called the multiple hypothesis testing issue. The issue here is that if you conduct enough statistical hypothesis, tests as is often the case, when the evaluations look at a lot of outcomes across multiple sites, and across multiple subgroups, we increase the chance of identifying one or more statistically significant when the true impact is 0, in response to that problem, two of the studies use approach to limit that issue.

The BOND study and the YTD study both used an approach called developing confirmatory hypotheses where only a few of the many hypothesis can test that were conducted were listed as confirmatory and

were focused on for policy implications. And the others were considered exploratory. In addition to the BOND study, I used the statistical correction to adjust the significance levels to account for the fact that there were multiple hypotheses tested. And we thought that the approach used in BOND and in the YTD are worth replicating your studies. Final point on the evaluations and data is that most of the studies took place in multiple sites and most, most of the evaluations that had similar treatments across sites pooled the data to come up with an overall impact. But some of the evaluations had very diverse, either treatments or groups of people across sites. And in those cases, the estimation was done separately for each site. Next slide, please. Beyond the impact evaluation. Well, our focus in the chapter is on impact evaluations, but there are other types of evaluation we just want to mention them here.

In addition to implementation You have impact evaluation, you have implementation or process studies, participation analysis and cost benefit analysis. Implementation studies are important to find out how the specific demonstration was implemented what the comparison, or control group got and whether people understood the treatment, and whether the demonstration was implemented with fidelity, these can be very important. And most of the studies did include that. Participation analysis looks at the relationship between the intervention, and which people participate. Finally cost benefit analysis is very important for making a decision on whether program is worth implementing overall cost benefit analysis, assigns monetary values to all the benefits and then compare them to the cost the treatment. Last slide please. In sum, Social Security Administration's decades of demonstrations provide quite rigorous evidence on the impacts of various interventions aim to improve the labor market and well-being of individuals with disabilities and to improve the application process. These evaluations provide useful models and lessons for future evaluation research, in terms of ensuring high quality evidence to inform public cloud public policy. And we would like to encourage social security to continue this admirable record. Thank you. Thank you, Burt. Now I will turn it over to Jesse Rothstein for some comments. Thank you very much for having me. You can put up my next slide.

So, I really enjoyed reading this paper. It was a great overview of, and kind of a catalog of, all the different evaluations and demonstration studies that SSA has done. I copied this table out of the paper just to show people the comprehensive list here. And I found it really useful and kind of thinking through how they all related to each other. Uh, as a discussant, I have the luxury of not being comprehensive and just picking on the pieces that I find interesting. So I'm going to do that. I'll divide it up into two parts. One, kind of talking about a couple of the aspects of the design that this paper talked about that, I think deserve a little bit more attention in the design, the future studies. And then second, talking about how there there's some interplay between the design of the study and how the how the results will be used for policy purposes. Next slide. So I'll start with a couple of design aspects. One: the paper does a good job in the presentation. It's a good job of talking about how it's really important in each of these evaluations to include a power calculation. And to make sure that before you start the study and you know exactly how much precision you're going to have. But it's also important to treat that not just as a technical part of the of the analysis report. But is it a part of the design of the study in the first place. The useful thing about power calculation is not that you do the calculation, but that you think beforehand about how big the effect needs to be.

In order to be what in medicine, they call clinically significant. And then make sure that your study is powered to be able to identify that so you need to before you design the study, decide how big or small the effects need to be didn't matter. And then, after the study comparable results to that often, in these

evaluations, it ends up working in somewhat the reverse. We have a budget, we figure out how many observations we're going to be able to pay for. And then we do a power calculation and figure out how big how big of an effect will be able to detect. Without ever really going back and saying, well, is that a big enough effect or is that a small enough effect that this will be a useful result? So, I think it's important to build in the kind of policy part of that into what looks like a technical piece. The 2nd issue is, is on the representativeness of the samples the authors talk about population representativeness. And I think that's an important issue to think about. But in a lot of these studies, partly by nature of the way that the kind of. Legal authority under, which they're conducted, you're really relying on people to volunteer for the program. And if they know what the program is going to be, that they're volunteering to potentially be randomized into. You have to worry a lot that the people who volunteer are the ones who expect to gain from this new program. And that makes them pretty non representative even if they look demographically representative on on traditional measures. And but that really limits our ability to generalize the results of the studies to thinking about what would happen if we rolled out these new policies to alert to the whole population, not just the volunteers.

A suggestion for the future is to think more about building study of that question into the design of the studies. For example, you could randomly assign some people to be eligible for an extra incentive that they participate. The bigger the incentive to participate, the more people you expect to join even if the specific treatment isn't, isn't all that useful for them and the more representative you'd expect the participants to be. And if you can measure how the treatment affect varies with the incentive to the people face to participate. You can then learn something about how important that self selection problem is. A third issue, this is a really technical point on missing data. The paper talked about how, when there's an item missing, the studies have usually done mean imputation. I think it would be a worth, at least exploring multiple imputation strategies to try to understand how. How sensitive the results are to the imputation, but that's a fairly technical detail. A more substantive detail is about is about site effects. These demonstrations typically take place in a few different, a few sites a few different locations. Um, usually a small number, so we don't have enough observations to really study the variation across the sites. But if you look at other kinds of evaluations that have been done by other agencies, and other settings, we often find that the site effects are really large that a given program has very different effect on some locations than in other locations often in ways that weren't anticipated before we before we set up the evaluation.

That's important to understand and I think when we only have 3 or 4 sites, we really limit our ability to understand that. But I think it's really useful for thinking about what the, what the results of the study will tell us for, for broader questions. About how to implement policy if a program is works. Moderately well, on average, that's one thing that if it works extremely well in some places and not at all well, in other places that's going to have very different implications for what would happen if we roll it out more generally. Especially if it ends up working especially well in places that are not random but are that have particular characteristics that may not be all that common in the broader world? For example, if you have implementation partners, and some of them are really hotshot and some of them are only okay and the program only works when you have a really hotshot implementation partners.

That tells you something about how well it's going to work when you roll it out more generally. And I think that's, that's something we need to understand better. It's hard. There's a reason why we only use a few sites, but I think it's, it's something that that is worth at least thinking hard about it gets added a

distinction that was that was prominent in the paper about efficacy studies versus efficiency studies that I'll talk about more in the rest of my comments. But how, if we're interested in just understanding whether the program could work, then then maybe understanding the site effects isn't so important. But if we're interested in, whether it will work on a large scale. Understanding the distribution of these site effects and and how they will, how they relate in the sample sites that are in this study to how they, what they would look like in the broader population is really important. And with that, let me ask for the next slide. Oh, and I apologize--I forgot a slide.

That's okay -- and one more slide. I want to talk now about the interplay between the design and the evaluation, and the usage of the results and policy. The paper talks a lot about the distinction I just mentioned between efficacy and effectiveness where efficacy is about testing the optimum implementation of the intervention on a small scale and effectiveness is about considering the program in a real world setting often at a large scale. I think it's useful to step back from the design of evaluations and think about what are the questions we would like to answer from a policy perspective. And I'm going to, I think of them as kind of 4 steps. 1, is we want to know whether it to you and mechanism operates. If we have a theory about how to get more people to work, we want to understand whether that theory is. Right that is a problem that is stopping people from working. Second, we want to know whether we can design a program to activate that mechanism. If the mechanism is real, is there any real world program we could implement that? We did we get at it. Third, assuming we can design a program to activate that mechanism. There may be many possible programs. And what's the best, most effective program among those that that do. And then fourth: once we've implemented a program, we want to know whether it's successful. Where that could have to do with the theory of the program, but more often it has to do with whether we connect, whether actually scales successfully. Or whether it winds up being a different program, when it's actually an operation than it was, when it was in the on the planning stage, and those are related to this efficacy versus effectiveness. But the way that the studies end up changing into policy doesn't always line up with that perfectly well, so I thought I'd take an example with the Ticket to Work evaluation.

This is a program that gave SSI and SSDI participants vouchers that they could use for training employment services and the real question and the evaluation was, did this get people more people to work. The mechanism we're interested in here is. Are there skills gaps that are stopping people from being employed? Are there things that employment services can help people match the jobs? Are those the barriers that are stopping people from getting jobs? And we'd really like to know whether that's true. Independent of whether any program, any particular program is successful at giving people those skills or or providing people services. Then, once we, if we knew that that mechanism, there's an operation, we then want to know whether we can design programs that overcome those barriers. And what program features are successful at getting people to actually show up for the training and getting them over to the hurdles they face. And then, once we designed a program that can do that at a small scale, we want to know whether it can continue to work at scale. How does this actually work in practice? Well, next slide please. We often jump straight to testing effectiveness.

We have a theory in mind about a barrier that might be there. We then design a program that we think gets at that theory. And we go straight to studying that is a very specific program as a program without ever, really studying. The question of is this barrier a live barrier with the same tools. I am borrowing an idea here from Jeff Kling and Coauthors about what they call mechanism experiments where you might

design an experiment. That's not really about testing a specific program that you would ever roll out on a larger scale. It's designed, the sole goal is testing whether a mechanism operates. And you might use a program that, that you would never scale. What we end up doing a practice as we test our program or very specific program on a fairly small scale. You might design that program to test the mechanism, even though, you know that you're not going to. To use that program on a large scale. What we end up doing a practice as we test our program or very specific program on a fairly small scale. We never really tested at a large scale, but we act as if the results apply at the large scale, when we get our results back, we act as if that says the program works or doesn't work.

Rather than just a proof of concept that it can work or not. If the program fails, we often interpret that as saying that the mechanism doesn't operate. But, in fact, the mechanism, might well operate, and we just have a bad program to get at it. And then if it succeeds, we rarely get around to testing it again at a large scale. If the once the program works, we say, okay, we're done we're going to just roll this program out or a program like it. And doing the, the large scale, uh, effectiveness studies is as much rarer, partly because they're very expensive to do. But we just assume that that works and I will stop there. This is a great paper. It gave me a lot of useful thoughts and I think it was it's a great way of tee off this conference. Great. Thank you so much Jesse for those thoughtful comments. Now I'm going to turn it over to Jack Smalligan again and we'll take some time to share his thoughts.

Thank you and for context, I'm currently at the Urban Institute, but my remarks reflect 27 years at the office of management and budget much of the time. I was in the office that was responsible for the Social Security Administration. I agree with Jessie, that Burt and David did a very impressive paper, and a very thorough discussion of SSA's past demonstrations, evaluation methodologies they used, and the Um, evaluation types that should be considered for future demos. And, um, is there I want to pick up on Jesse's last slide where he raised the question of what questions are we trying to answer and really kind of focus on the agenda going forward? Um, or how we redesign kind of the demonstration authority. And also how we sort of think about this as a kind of a national demonstration authority. Not purely, um, focusing on the social security programs, um, but more generally how we can do demonstration projects to help people with disabilities. And so to do that the focus for the demonstration authority should be redesigned to be much broader and as Burt and David described the current unit of analysis as individuals receiving, or potentially receiving desire, both benefits.

And if we view this as more of a national demonstration, if they're already, we'd recognize that there's many people who identify as having a disability who have no attachment to either SSDI or SSI. And In that, um, there are a lot of other programs that are serving that population, but if Generally much less funds for demonstrations and much more limited, legal authorities perhaps with the exception of CMS and, um. Consider for instance, the Department of Education, vocational rehabilitation system, we receive really, very few rigorous evaluations. That is random assignment type evaluations within the VR system and yet we're spending billions of dollars there. And, um, and yet and do not have the evidence based to know what types of VR services are most effective. And considered the National Workers Compensation program Here the most interesting experiment was what Washington state did with the Centers for Occupational Health and Education. But the impetus for that was entirely at the state level and at the at the academic level, the impetus was not at the federal level. And yet we learned a great deal from that, um, that effort and, um. And consider state and private sector, um, short term disability and insurance benefits. We have a very limited U.S. evaluation research experience , even though at an

international level, there's a lot of experimentation that's going on around short term disability benefits. Um. This current focus with many evaluations on returning to work for existing beneficiaries and with a priority on achieving budget savings for SSDI and SSI um, is missing an opportunity and with a. Broader charter, we could better test any evaluation interventions or programs intervene far earlier with, at risk individuals who have really no, um, connection to security programs or even an awareness of the programs.

Department of Labor's Office of Disability and Employment Policy, and SSA have made a start on this with the RETAIN Demonstration that is retaining employment and talent after injury and illness. And then there is a lot of other proposals, from people like Jennifer Christian, David Stapleton. David Mann Yoni Ben Shalom and many others. And looking ahead, there's a great deal of interest in government paid medical, leave as part of comprehensive paid, parental care giving and own medical leave. We see strong support at both the state in the federal level for that kind of a program. When we have a medical leave benefit in state program, now it's frequently for as much as 12 weeks of own medical leave and how should we intervene when a worker exhausts that medical leave benefits and they're still not capable of returning to work. These programs have a readily identifiable target population that's at risk and should be the scope of more rigorous evaluations for finding ways to intervene and help them stay in the workforce. My second recommendation is that, in terms of specific SSA demonstrations, we need to examine SSA's own internal eligibility determination process. So how do we design process evaluations that are not evaluating a new intervention but are evaluating SSA's own internal disability determination, steps or processes? And as past research, as Burt and David described where there was a focus on the determination process were mainly proof of concept type evaluations. And in some cases, and narrow targeted intervention, such as for people who are homeless.

All kind of useful effort, but far too limited in terms of what SSA has needed, given it the serious problems we have with SSA's current process particularly the years. It can take someone to receive an ultimate disability of determination. Consider, for example, for about 20 years SSA didn't perform reconsideration reviews at the state level in about 20% of the country. Recently, SSA began to re-instate those reviews, even though the evidence base was limited and various experts disagreed on fundamental aspects of the value of that stage of the process. States perform these initial determination. and they performed the reconsideration reviews. And some other programs, we leverage state level, variation and program delivery to learn more, but we don't do this in SSA. And it's a missed opportunity. We could test more or less intensive reconsideration reviews and compare their effectiveness. So, we know SSA's protracted process that takes years to have a person receive a disability determination has serious problems, um, research by Maestes, Mullen and Strand shows that a very delayed decision leads to decay and, uh, the applicants work capacity. In other words, SSA's today's current process that causes some people to wait for several years for a decision functions in effect as an intervention with adverse outcomes for those denied applicants.

That those findings are primarily an indictment of the protracted hearings level stage of the review. Could we improve the earlier stages? So fewer applicants need to go before an ALJ. Commissioner Barnhart tested an alternative, but the test was suspended under Commissioner Astrue before we had the full results. How do we design re, how do we redesign the process to function better? And how do we evaluate those efforts? If we invest more in making better decisions earlier, how do we evaluate what is a better decision? Is another area requires Congress to reimagined the role of SSA's current this

is demonstration authority and to redesign a broader mandate for both SSA and for other agencies. Thank you, thank you so much Jack for those comments we are going to take about a 5 minute break and resume with the next session in this panel at 10 till the top of the hour.

10:50 – 11:25 a.m. EDT: Use of Demonstrations

Presenters: Austin Nichols, Abt Associates
Robert Weathers, U.S. Social Security Administration
Discussants: Jonah Gelbach, University of California, Berkeley
Elizabeth Curda, U.S. Government Accountability Office

Hello, everyone, we're about to get started panel of the State of the Science Meeting on Lessons Learned from SSA Demonstrations. This is the second presentation and will focus on the use of demonstrations. Robert Weathers and Austin. Nichols have written a delightful paper on this topic. And Bob will share with you some insight from that paper over the course of, about the next 15 minutes after which we'll have two discussants Jonah Gelbach and Elizabeth Curtis, they'll share their thoughts on this paper. Then at what we're guessing will be about 11:25 Eastern, we'll see, we will have general discussion and Q&A. So, on that point, you are free to place any questions that you have, in the Q&A box on Webex. And for now, I will turn it over to Bob Weather's on the use of demonstrations. Thank you Laura, it's been a pleasure to co-author this chapter with Austin Nichols. These are our views, and not necessarily the views of the Social Security Administration, or the federal government. Next slide. Our chapter covers 5 broad topics: 1) when to use a demonstration, 2) when to evaluate, 3) improving demonstration and evaluation design, 4) Improving data use both surveys and administrative data, and 5)maximizing the use of demonstration findings and building a stronger evidence base.

A demonstration is a trial program or a policy change, and we refer to such changes as an intervention. Many of us SSA's, demonstrations are applied to a random subset of the population referred to as a treatment group, and another randomly selected group that receives business as usual serves as the control group. This is referred to as a randomized control trial. And when they're practical and appropriate, they're a rigorous way of identifying causal impacts and intervention. Depending on policy constraints, it's possible to assess a program or a policy. In other ways. At one end of the spectrum, an initial pilot project might be used to work through implementation and operational issues with a new program and to assess its feasibility. At the other end of the spectrum, simply rolling out the intervention. On a national scale, using randomly staggered roll out. Might be appropriate to provide a basis for program evaluation. For a variety of reasons, some evaluations are more practical to do outside of a demonstration context. For example, the effects of interest. Might be too small to identify using a demonstration, or it might be difficult to approximate conditions of an ongoing national program. Program entry effects are an example of where these factors make it impractical to use the demonstration. SSA has done a tremendous job in implementing and evaluating programs through demonstrations as well as using other methods. But there are opportunities to do more.

Next slide. One opportunity is to design projects and form a wider array of policy options. For example, one work disincentive within the disability program is a complete elimination of the monthly benefit, check, much a person earns and earnings amount about the substantial gainful activity level. This loss and benefits is referred to as the cash cliff. And it's a substantial additional tax on earnings. Policies that

gradually reduce benefits as earnings increase referred to as benefit offsets. Are aimed at reducing the work disincentive, but they maintain some additional tax on earnings. There are a variety of ways to gradually reduce benefits. Some may be more effective at encouraging work than others. The benefit offset demonstrations to date focused on a 1 dollar for 2 dollar benefit, offset with some variation and the amount of earnings. Or the time period where the benefit offset would begin. We suggest expanding the range of potential benefit offsets, for example, a demonstration that completely eliminates the extra tax by allowing beneficiaries to work and maintain their entire benefit payment. This would test the extent to which the cash cliff is a work disincentive.

However exploring a range of benefit offset policies would be desirable in obtaining a more complete picture. Factorial designs as described in Barnow and Greenberg are one option to do this. Another opportunity is to make greater use of theoretical models. We identify several advantages of using these models 1st, they may identify counterintuitive outcomes. For example, a simple economic labor supply model describes unexpected effects from a benefit offset. The counterintuitive effect is due to those who have their benefits suspended due to work. Activity under existing rules--under the benefit under the benefit offset the model illustrates that they will receive a partial benefit under the benefit offset. And some may even reduce their work activity because of the extra income they receive from the benefit offset. We found this to be the case and the benefit offset pilot demonstration. second, a well specified theoretical model may be estimated and validated. Using data from a demonstration project and then use to simulate policy responses for other policy changes. This approach has been used by others, for example, for example, Petra, Todd and Ken Waltman, use results from the school subsidy program. To estimate, and validate lifecycle model fertility decisions, and the number of years of education, they then use the model to simulate the effects of alternative subsidies on fertility and education outcomes. A similar approach could be used by an SSA demonstrations.

The use of falsifiable logic model, or another opportunity. SSA has done a commendable job of incorporating logic models and the demonstration projects. Almost all the design reports identified, expect expected inputs. Outcomes and impacts from an intervention. But do so qualitatively Epstein and propose the use of a falsifiable logic models that involve specifying in greater detail the expected inputs and intermediate outcomes in the logic model that are expected to produce quantitative outcomes. If the inputs and outcomes are not achieved and that we would have less confidence that the impacts are theoretically achievable. They propose this as a means for identifying pilot projects that might be ready for a more rigorous evaluation and minimizing the costs potentially expensive demonstrations that show small effects as is the case in some of the SSA demonstrations. But they're also useful for providing quicker results of the potential effects of a rigorous demonstration. We see an opportunity for SSA to examine variation or heterogeneity in average treatment effects. Examining differences in average treatment effects can answer the questions about for whom the policy is effective.

Differences and impacts by specific subgroups, for example, by age impairment type or other characteristics is one way to address the “for whom” question. And this will be covered later today by Till Von Wachter. Interventions might have different effects among those at different points in the distribution such as in earnings or income distribution. Thus the standard approach of measuring mean impact to miss meaningful effects for an intervention. A good example is a paper by Bitler Gelbach and Hoynes who examine such effects for a welfare reform experiment called Jobs First. This approach would be useful for SSA demonstrations where the effects may differ within the distribution, whether its

earnings, income, health, or other outcomes. Third, an intervention might not be used by all treatment group members and some control group members might access the intervention. In these instances we are interested in whether the intervention is effective for those who accessed it because they were assigned to the treatment group. By the effect of the treatment on the treated. the Oregon health insurance experiment is a great example of estimating such effects.

This experiment involved, offering, a randomly selected group, the opportunity to enroll in Medicaid and only 25% of the group took the offer Experiment estimated the effectiveness of being offered Medicaid effectiveness of using it. A similar approach is potentially important for many of SSA demonstrations, including Project Network. Promoting Readiness of Minors in SSI or PROMISE, and the Accelerated Benefits Demonstration. And there are also opportunities to enhance cost benefit analysis. Most of SSA demonstration projects, examine the effects of demonstration project from a Social Security Trust Fund perspective. BOND is an exception. And it includes a cost benefit analysis from society's perspective. This perspective is important many programs, produce a net cost to the government, but produce significant value to society. The approach that Abt used with BOND provides sufficient details on assumptions and methods and includes the sensitivity analysis. The estimates appear reasonable, given the impact estimates and the assumptions, but the assumptions have led some to criticize the approach. Another approach that is promoted by Nathan Hendren and co authors is also useful. At a high level their approach uses information to identify society's willingness to pay for a policy or program. An advancement of their approach is that it provides consistent method for assigning net value to various public programs and therefore allowing one to make comparisons across programs and policies. Next slide.

Evaluations of demonstration projects require good data. SSA demonstrations have relied on the collection of survey data to obtain the specific data needed for an evaluation. Among the concerns that survey data to stand out well, response rates and the quality of the responses. Importantly, low response rates are growing concern. The most recent national beneficiary survey, SSA conducted had about 65% response rate, well below the 80% standard identified by the Office of Management Budget. SSA demonstrations have also relied on administrative data to obtain information. Administrative data is generally available for all demonstration project participants. And there are fewer concerns about the quality of the day, compared to surveys. However, the data from other federal agencies would improve SSA demonstrations. Data from centers for Medicare and Medicaid services on expenditures. Data from the office of child support enforcement, quarterly wages. And data from the IRS on income, would significantly enhance. Information drawn from many of SSAs demonstrations. Combining survey, data, administrative data will be important for us to say to do in the future. Combining the data can be useful for an assessment of low survey response rates, or a potential selection bias. And the assessment of reliability of survey responses, for example, employment earnings could also be performed with administrative data David Wittenburg, Jeffrey Hemmeter and co authors have a nice paper that shows how this may be. Next slide.

Communications are another potential area for improvement, engaging stakeholders. Early and often is important. Stakeholders can be invaluable partners with the design and implementation of demonstration. SSA has begun to make greater use of technical evaluation panels, which is step in the right direction, but engaging other stakeholders such as the social security advisory board. And the disability community, through SSAs, national disability forums are two other ways that SSA has to

continue to improve demonstrations. Getting results out faster is another opportunity SSA might explore the use of shorter finding speed to highlight key results. Sooner. The office about evaluation sciences provides a model of short finding sprees. That can provide relatively quick in the private relatively quickly and that are accessible to a wide audience. The Department of Labor has a model of communicating results using email, and there are opportunities to make greater use of social media to communicate findings quickly and efficient.

Next slide building a stronger evidence base. First use the use of qualitative findings, SSA final reports that rely heavily on quantitative findings on impacts, which are the bottom line for most readers. But there are opportunities to make greater use of qualitative findings on the experiences of demonstration project participants to provide a more complete picture demonstration. Individual experiences can provide a powerful way of explaining quantitative results. Making data available through analysis is also important. By the final reports produced by demonstrations have been very informative. There's plenty of room for re-analysis of the demonstration replication and re-analysis are useful in adding additional credibility to the results. In addition extending the results through new analysis can provide policy makers with important information. Synthesizing results across demonstration projects. SSA has assembled a large body of high quality demonstrations. There are opportunities to synthesize results across all these projects. Today's conference is a great 1st step, but it took a long time to get here building out a separate in the future. Has the potential to improve the evidence. Policy makers may need. And further improve SSAs, approach to conducting demonstrations.

And next slide, so the key takeaways SSA is operating state of the art demonstrations at scale. And this is an enormous accomplishment. We identify opportunities to build on a strong foundation and leverage current opportunities. For example, the foundations for evidence based policy, making act at 2018, emphasizes the role that demonstration projects, and evaluations should play informing program. Our chapters focused on 3 questions. What research questions need to be answered, how should the questions be addressed and how can we maximize the use of answers that concludes our presentation. I'll pass it over to Laura and the discussants. Thank you. Thank you Bob. So now, our first discussant is Jonah Gelbach. I'll turn it over to him to reflect on this paper. Thanks very much. I'm Jonah Gelbach and I'm an economist and professor of law at Berkeley Law. I'm delighted to comment today on chapter 3 by Robert Weathers and Austin Nichols. This chapter provides a wide ranging assessment of. How to make the most out of the social security administrations demonstrations. I found the chapter both comprehensive and insightful.

I'll focus my comments on one observation Weathers and Nichols make. Quote that “past relevant demonstrations themselves might be subjected to a cost benefit analysis before launching a new demonstration." They suggest that e just such an analysis requires thinking broadly about the general goals of demonstrations ex ante. They're added value and how we might judge them ex post. That's also a quote. These are excellent suggestions because as weathers Nichols emphasize. Undertaking a new demonstration incurs substantial opportunity costs and not just those related to government funds. We'd like to know that the benefits justify the total cost. So can go to the 1st slide please a good example is the chapter's discussion of the impact of program parameters. On the composition of program participants and applicants. That is entry effects. Weathers and Nichols discuss SSAs consideration of whether such effects could be productively studied for SSDI.

Using a traditional randomized control trial, or RCT Because such a demonstration would need to target initial non participants would need to target a sample from the U. S. population. As a whole pitching, such a sample via a traditional demonstration would be expensive. Given that the population level SSDI participation rate is quite low. An expert review suggested that to have a reasonable shot at detecting effects of interesting magnitude. A demonstration would have to have something like 9 Million participants. Weathers and Nichols describe additional challenges raised by the expert review, and they note that in the BOND demonstration SSA chose to focus on questions related to reforms effects on current participants. That is only current participants. This discussion connects to a longstanding area of controversy amongst scholars studying causal effects Of social programs, how much can we learn from RCTs? Should we concentrate our evaluation resources in that domain? Larger scale, social demonstrations have a long history, including the negative income tax experiments of the 1970s. And the health insurance experiment of the 1970s, 1980s when state level welfare reforms are all the rage of the 1990s. Numerous RCT demonstrations occurred.

Now, the general argument for RCTs is familiar. They're supposed to balance differences and treated in control units so that researchers and policy makers may be confident observed outcome. Differences are due to an interventions, causal effects, rather than differences in confounders. For this reason smaller scale field have become more popular in recent years, especially in the area of development economics. From which economists Abhijit Banerjee, Esther Duflo, and Michael Kremer won the 2019 prize and economic science is in memory of Alfred Nobel. Explaining its choice of topic area the prize committee wrote and it's scientific background document that quote. The best way to draw precise conclusions about the true path from causes to effects is often to conduct a randomized controlled field trial note. The word best in that sentence. One frequently, here's the term gold standard, applied to our cities and what could be better than gold. Actually, there's a serious case that the gold standard contributed importantly to the scope of the Great Depression. This is a rhetorical point to be sure but it's a good reminder that apparently unassailable things can have their flaws. And the case against our isn't just rhetorical. Nobel oriented economics, Angus Deaton and philosopher of science.

Nancy Cartright wrote in a 2018 paper that quote any special status for RCT is unwarranted. The good statistical properties of RCTs hold only on average they point out rather than in any particular RCTs and is suffer precision related issues, which help explain why estimating SSDI entry affects whatever requirements. So many participants. Another issue is treatment effect heterogeneity. The idea that some people will respond more than others, or even in the opposite direction when facing change to policy parameters. The limits of RCTS related to treatment effect. Heterogeneity is a subject that has long been discussed in economics. Another Nobel laureate James Heckman pointed out in 1995 paper coauthored with with Jeffrey Smith. That RCTs do not identify the distribution of program gains unless conditional assumptions are maintained. That's important because distributional considerations often are quite important to policymakers to be sure the fact that RCTs have their limits in the presence of heterogeneous effects mean distributional knowledge is out of reach. That doesn't mean some circumspect and careful attention to underlying theoretical considerations.

Are warranted when considering RCT use the 2019 Nobel Prize committee itself, embraced the role of economic theory and policy design. At the risk immodesty, I'll point to my own work. Which Robert Weathers mentioned in his presentation. I coauthored with economists, Marianne Bitner and Hillary Hoynes, using data from Connecticut's jobs. 1st, welfare reform demonstration project. In studying Jobs

First we were able to connect predictions from basic labor supply theory to the ways in which a change in program parameters could be expected to operate across the earnings distribution of demonstration subjects. Our research was conducted entirely after the demonstration had finished. And it was possible, only because MDRCs data from the demonstration were available for use by researchers via a not too onerous process. This raises an additional point. The value of making demonstration data publicly available for further study. There are other criticisms of demonstrations.

For example, as Heckman and Smith pointed out in their 1995 paper, typical RCTs reveal only short run policy reform effect. Of course, the same is true about many non experimental evaluations. More generally as Banerjee and Duflo note the fact that RCTs have their problems doesn't mean that non experimental approaches are immune to those same problems. So, what lessons can we draw from this discussion? 1st well designed well executed RCTs solve a particular class of statistical problem. They balance treated in control units on the distribution of confounding effects. On average within experiments. that allows someone with the data in hand to estimate some kinds of parameters that may be of policy interest. But 2nd, even perfectly implemented RCTs don't allow us to answer every question of interest either because some questions such as entry effects. Are by their very nature difficult or expensive to study at all with RCTs, or because of the extent in nature of treatment effect heterogeneity. So, whether RCTs are better than alternatives in any given context, depends. It depends on the questions that are of interest on the policy reform options.

On the initially unknown distribution of people's responses to policy reforms under consideration and on what will be done with the information obtained from the demonstration. Of course, whether RCTs are worth doing also depends on the alternative. We should always ask compared to what. next slide please. There is a long history of non experimental estimation in the social sciences. Both structural and reduced form econometric methods have developed in important part for the purpose of answering the kinds of questions that RCTs would answer if they exist. These points are not unknown to the discussion SSA demonstrations of the discussion of entry effects That Weathers and Nichols offer at pages 12 through 14 of their chapter illustrates. They cite an SSA funded, RAND paper by Nicole Maestes and co-authors Mullen and Zamarro Titled "Research designs for estimating, induced entry into the SSDI program, resulting from the benefit offset," which describes 2 RCT alternatives, stated preferences and structural estimation using variation from policy changes.

These authors considered, but rejected alternative approaches. Including more complex, structural models. I suggest here that even where RCTs are feasible to design and administer at manageable cost, it is not obvious that they are always the best choice. One way to look at this issue is to recognize that the choice to use an RCT to study a question is itself a policy choice. The internal logic of the contention that are RCTs are necessary for better policy study requires randomizing whether RCTs are used to study questions. That seems unlikely. what we have available is the considered ex ante judgment of experts. SSA ought to use that resource liberally. I have one final suggestion next slide. Please. SSA ought to invest in making its data more available to researchers operating outside, either the agency itself, or its contracted parties. There are lots of highly skilled researchers who want to study questions that are, or would be of interest to policymakers but who aren't able to do so, because they can't get data. SSA can radically increase the amount of available research knowledge. If it found ways to make existing administrative data, more publicly accessible.

Of course, there are privacy considerations and program operations must continue without interruption, but perhaps SSA should consider whether the next demonstration is likely to lead to information is valuable if we might. Were to spend some of its resources figuring out how to productively share data for a wider study. In sum, I applaud Weathers in Nichols general suggestion that more thought should be given to whether particular demonstrations are worth the expense and time it will take to conduct them. There are alternatives, including non experimental study in particular settings and wider data access in general. It's due SSA's credit, that the agency has commissioned this volume. And it will be to all of our benefit if the agency follows these authors suggestions. Thank you, thank you, Jonah. I appreciate you taking the time to read and consider the paper and share your thoughts. And now I'll turn it over to Elizabeth Curda for her thoughts on this.

Good morning, and thank you for the invitation to discuss Robert weathers and Austin Nichol's paper on a very important topic improving the use of demonstration projects. I'm Elizabeth Curda, a director at the Government Accountability Office and I typically oversee our reports on disability programs and issues. GAO provides audit and evaluation services to Congress, regarding federal agencies and programs. Because of this GAO has a body of work that speaks to many of the issues the authors raised. Our work covers a wide array of activity is conducted by the federal government, such as demonstration projects, impact, evaluations and program magic.

Next slide over the last 20 years has carried out many demonstration programs and spent hundreds of millions of dollars doing it. I really appreciate the author's focus on improving the use of demonstrations because the cost and consequences of a demonstration not being useful. A well designed demonstration project that produces useful results can influence policy and potentially affect the lives of millions of disability beneficiaries. Chapter 3 suggests an array of possible improvements to enhance the effectiveness of SSA demonstration projects. These suggestions seem to fall into 3 main buckets, methodological process and communication improvements. I'm going to focus my remarks today on areas where the authors suggestions for improvement dovetail with prior GAO work. Many of these suggestions overlap with recommendations GAO has made in the past as well as best practices. We have identified. Both for demonstration projects specifically. As well, as for project management and federal, internal control standards that federal agencies should be using to manage programs more broadly. Next slide. The authors make a number of important points with respect to how demonstration projects can be most effectively designed.

The 1 that stands out as a key design element is the suggestion to employ logic models in order to summarize assumptions, outcome, pathways and causal mechanisms. In the past GAO has discussed in several reports that logic models can provide a framework that links the intervention to the goals and outcomes desired and as recommended their use in developing new programs as well as evaluations of programs. The authors also suggest demonstrations can and should test multiple intervention options and causal channels through multi stage, multi arm or factorial design of interventions and experimental evaluations, leveraging a large and costly demonstration project to test several policy options and explore multiple causal methods. Could be a good way to increase the bang for the back of a given demonstration. However, doing so increases the complexity of a demonstration and requires careful, design to be effective. GAO's work on designing evaluation stresses the need to be extremely clear about the evaluation questions at each phase of a project, and what is being assessed --processes versus outcomes versus impacts of alternative interventions, for example, and to select appropriate measures

and criteria for success at each stage of a programs. Implementation. GAO has also stressed the need to assess an intervention effects compared to a counterfactual of what would have occurred without the intervention.

For instance, in our 2008 report on demonstration projects, we found that some demonstrations did not assess the project's effects compared to what would have happened in its absence. Further there we noted that planning for the evaluation has to be part of the demonstrations project design. As part of that report, we recommended the SSA implement clear written policies and procedures that are consistent with standard research processes and federal internal control standards. As a result, SSA developed a demonstration product project guidebook, which outlined the agencies policies, procedures, and mechanisms for managing and operating its demonstration projects. To the extent SSA adopts any other recommendations from this panel, incorporating them into the guidebook with better ensure future evaluators, consider their use.

A key aspect of federal internal control standards that apply to demonstration projects is ensuring that high quality data are used and data needs are clearly identified at the beginning of a project. in a recent report on HHS demonstration projects. We also pointed out that it is critical that baseline data are collected in comparison groups are established prior to the intervention. Another key principle that I think should be highlighted is the need to identify potential risks to the success of the demonstration project in advance, identifying risks as well as being prepared to analyze and respond to the risks is another principle for internal control and encompasses at a high level some of the methodological considerations the authors highlight. This includes identifying and documenting trade offs in scoping the demonstration project, identifying ways that the proposed methods or timeframes may fail to meet the needs of policy makers. And clearly understanding potential sources of bias in the analysis. Next slide. In the chapter, the author's stress, the need to build a better evidence base by, among other things, using qualitative information to provide context and explore causal mechanisms.

When reporting results. Taking that a step further, considering participant voices in the planning. And design of an intervention could yield to new insights and identify potential risks to the success of a demonstration, for example, in a 2010 forum on disability reform held by GAO stakeholders, including those, with a participant perspective noted that new benefits services and programs need to be carefully structured to avoid unintended consequences and that the costs and benefits to stakeholders, including participants must be considered in the program design. Another process improvement I would add is to take steps to ensure transparency about changes to the demonstration along the way. GAO has found that changes during a demonstration can cause problems and affect the quality of the evaluation. Changes to the design of the demonstration. The sample. Related policies that may affect participants, et cetera should be documented along the way along with plans. for how those changes will be handled in the evaluation. Next slide the author stress, the need to involve stakeholders early and often as well as to disseminate interim results to key stakeholders.

This is standard operating procedure at GAO as our internal control standards, require appropriate internal and external communication for example, at a bare minimum. We typically gather stakeholders at project initiation then to review and comment on the proposed project design then to approve the final project design. To discuss preliminary results and assess the evidence. To develop consensus around the meaning of findings, and to review and approve the key messages that the evidence

supports and after all of that we include project stakeholders and reviewing report drafts. while this type of collaborative process takes time you find it typically leads to more robust results greater acceptance of our findings and fewer surprises and rework at the end of the day. The authors also state that demonstration findings should be leveraged to affect policy through good communication. This is critical, and it's not a universal practice. GAO has recommended as recently as 2018. that agencies provide rigorous final evaluation reports and publicly released findings of demonstration projects. In addition in our 2000 report on SSA demonstrations, we recommended that SSA established a formal final report process that includes a complete post mortem on the project and identifies outcomes limitations and policy options for Congress SSA implemented this recommendation through its annual reports to Congress. on its demonstrations and papers published in academic journals and online. However, I would point out that communication can be continuously improved. In our experience at GAO, it's necessary to communicate policy results.

To Congress, in a manner far different than one would communicate in academic journals. A key aspect of this is developing high quality communications that is geared toward the intended audience, which includes government agencies Congress as well as the general public. This is challenging. Because we know at GAO , however, are working to make the findings of demonstrations, more reader, friendly and accessible to key. Audiences would go a long way to improving policy makers, ability to leverage demonstration results into action. And finally going beyond the focus on individual demonstration projects, GAO has previously identified more than 40 programs managed by 5 different federal agencies that provide a patchwork of employment support for people with disabilities. We reported in 2012 that these programs lacked a unified vision strategy, or set of goals to guide their outcomes. GAO has been recommending since 2012 that OMB work with federal agencies to coordinate the development of a set of unifying government, wide goals for employment of people with disabilities. Such an effort could provide much needed focus and impetus for designing demonstration projects that align with federal employment goals. It could also help pave the way toward taking greater advantage of the wealth of federal data that is collected by different federal agency agencies, but requires Herculean efforts by agencies and researchers to share and use in these important evaluation efforts. Thank you. And I look forward to your questions.

Thank you Liz that was so interesting to me and I think that your penultimate point in particular is a great transition to this discussion, right? You were pointing out how dissemination efforts are really essential to ensuring that the findings from demonstrations. Reach various target audiences and that's what this whole meeting is about. So thank you for that. I'd like to start with the Q&A, by asking any of the, the presenters or authors. If they would like to share. Comments in response to the discussant remarks. So that would be from the 1st paper Burt and Dave, whether either of you want to say anything in response to the discussants, or or likewise from, from this, most, recent panel, Austin or Bob. Are you, do you have a response.

Well, this is Burt, I just want to say, I appreciated the comments from both discussants. I think they were very helpful and I would have liked to have had more time to talk about equity and efficiency trials as Jesse did. That was really good, but I agreed with all the comments and I think they would be very helpful. Bob or Austin do you have any Observations in response to the comments, you would just say that there are lots of interesting ideas being discussed here. Some of these. The past demonstrations that are discussed in in volume, and that are in the presentation as well are mandated by Congress. So

they're design features that are not easily accessible to SSA in some cases. But, obviously, we would like to put those ideas out there anyway and to just and to sort of map out the space of demonstrations that could be conducted in the future or non demonstration evaluations. So that's an interesting point. And it's 1 that actually came in relates to a question that came in from the audience about the linking of Medicaid and SSI and SSDI Or, and so you're making the observation that. Um, Congress has a strong hold on permission to a certain extent.

Can you or someone talk about how. one might go about changing the law needed not only for kind of mandatory participation. Of claimants, or beneficiaries in demonstrations, but then also, that observation about the, the linking of multiple. Programs and how you might understand the kind of the. Independent and collective effects of multiple social programs operating. I may be Bob since you are an insider in the agency. That is a question for you. Sure, I'll, I'll take a shot at that. I think 1 opportunity is through the Foundations for Evidence Based Policymaking Act there's. An effort to really have federal agencies work, more collaboratively together, not just in terms of data sharing, but in terms of service delivery and I think that that's also a priority of the administration. So, I think that in the coming years, we're going to see more collaborative efforts across federal agencies. And I will point out that SSA has worked with DOL and has worked with the Department of Education on, on some recent efforts that I think are a step in that direction.

Great. So, mentioning the, uh, the Administration's priorities, we also know that the President recently issued an executive order on equity and I would like to hear from any and all of you, how you think demonstrations in the future can ensure an attention to equity I guess both. With respect to the interventions as designed and tested, and also with respect to the approach to evaluating them. Hello. Laura, I have some comments on that I, I've given you a lot of thought, because our program at George Washington University is emphasizing equity as well. Equity is not a simple topic, but a lot of the different aspects can and have been addressed in the social security demonstrations and others. So, let me just talk about. Some of the aspects of equity 1 has to do with access. And the question is, are people in different groups receiving equitable access, maybe it's equal, or maybe it's based on priority and that can be accomplished in several ways. One, to do a quantitatively is the participation analysis that we've mentioned where you can actually try and assess quantitative how to different groups, or offered the opportunity. How how do they participate? And that can suggest different ways. You can improve access to the programs. Secondly, impact evaluations are really amenable to looking at equity and there's a couple tricky aspects.

There 1, is that the subgroup analysis that we discussed does look can easily look specifically at equity of impact. If you have a sufficiently large sample size to look at. How does the program affect different people and are the impacts. Similar or different across groups, but equal impacts isn't necessarily the only thing this is one area where it does get somewhat tricky, because some groups may start out more disadvantaged than others. And part of the equity concern may be closing the gap. So, you'd want to look at more than just the magnitude of the impact possibly. And then finally, the process studies can be very important for trying to understand how programs are implemented and whether groups are affected differently. I think this has come to light during the pandemic. Especially where we know that some groups, especially African Americans have that more reluctant than others to participate in the vaccine, because of past history. So the process analysis can help get that.

Those are some of the ways that you can address equity. I appreciate that what thoughts do others have on that topic. I think I just added we need better are going to baseline information about the applicants for benefits to understand the characteristics, um, in terms of in a racial and, um, diversity and income. And just because we don't know some fundamental things in terms of the rate at which people are being denied benefits and whether that varies across states. And, um, what kind of characteristics would explain, um, those different allowance and denial rate. So before we actually do experiments, I think we need a better baseline for how equitable the program is. Um, as it's functioning today. one just kind of statistical thought on top of the kind of very good point that other people have made.

The a lot depends on how exactly we frame the question. If we frame the question as if there are say, 2 subgroups. As is the effect of the program equally good for the 2 groups that that creates a situation where you need quadruple the sample size of what you would need to do to get the average effect. And that's often prohibitive until we end up. Not really being able to study that very well, but if you frame the question, as does this program benefit both groups. Without really needing to be able to tell whether it's the same benefit, just making sure that it is beneficial that only requires double the sample size and it's a little more feasible. And so, so, I think for the equity purposes, ideally, we'd like to know if it's equal but I think it's enough to know that the program is beneficial to both groups. For many purposes yeah, that's a really helpful observation and I think to what the, the notion of closing gaps, right? So, if people don't start on equal footing, but they end on. Closer to equal footing that could be a. Um, a move forward as well. I just wanted to say related to that, or I dip into another question other thoughts on this yeah, please. Sorry related to Jesse's point about bringing the question. I think another thing that should happen in future. Evaluations and demonstrations is a more diverse group of researchers should be involved in those projects from the very earliest stages.

So, a lot of the, you know, when we talk about our logic model or using a logic model, that's embedding a lot of the design. Into a set of causal relationships that it sort of pinned down where the evaluation goes from there and having. Folks from different walks of life and researchers of color involved in those early stages could change the nature of the questions that are being asked. I just wanted to to echo that point. I have never designed or conducted and evaluation myself, although I've worked with. Data exposed, um, I can't imagine all of the many different issues that arise and thinking about how. On how a social security. Demonstrations particularly one related to SSDI would. Affect different people, and I just think that I'm without meaning in any way to besmirch anybody's choices here. I'm looking at a screen full of white people right now and, you know, we're talking after. The fact, and tell him, but I do think that echo Austin's point. I, I don't know anything at all about who is designing and conducting demonstrations already, but I do think that ensuring that there's a broad cross section, particularly along racial and ethnicity grounds. It seems to me a, a very important 1 to take seriously. And I just wanted to add that that GAO is increasingly being asked to look at equity considerations and the programs that we evaluate. Um, and I can't speak to necessarily other agencies, but I know 1 challenge in examining racial equity issues. That SSA has some limitations in the data.

That exists to be able to do that sort of easily. There are some sort of complicated data matching things that you can do, but they're less precise and it has to do with just a change in the way SSA started collecting or stopped collecting this data at some point. So, there are some limitations there as to what can do. Thank you for your thoughts on that I'm going to move to another question, which is to. You know, for your papers, you focused and delved very deeply into the rich history of SSA demonstrations.

I'm going to ask you to step out of that for a second, and think about what non SSA experiment. Offers a model for SSA is there something that we should be learning from, from elsewhere in social policy evaluation that allows us to translate some useful advice for SSA in the future? So. Let's start with maybe with the authors from, from chapter 2. Burt and Dave, do either of you have a top pick for your favorite non SSA experiment that has insights for SSA. I'm not sure the. Specific experiment that, um. There's been a lot of work on what the effects, our own other family members in various experiments Most SSA experiments as best. I can tell focus almost entirely. On the people receiving the DI or SSI but Say, very little about whether a particular treatment causes, say, a spouse to go to work or stop working what the effects are on children, that sort of thing. And I think other. Um, demonstrations provide good models for doing that. That I Austin or Bob from. Oh, please, go ahead.

This is Burt. I just wanted to add that. I found some of the work that the administration for children and families has done be kind of interesting. They sometimes start with very big, but difficult questions. Like, how can we improve self sufficiency and then. They make an award and the actually Abt has done some of these as well as other large for profit non profit firms. First currently, there's the PACE e study which the 1st, 3 years or so were devoted to coming up with. What would the design be? What's the right way to attack the problem? And I don't know that Social Security hasn't done that but and it's a luxury. Sometimes. You don't have. Especially if Congress is telling you to give you answers in 3 years. But I did think there is something to that. Where the contractor had 3 years to come up with okay. What's the best way to address this? What do we already know? What should we be testing? So I'd like to see more of that and not just SSA , but other agencies. I will say that the benefit. The benefits of offset demonstrations I should say, and at least a 7 year run up in terms of design. So, there was a long period of time where various designs were routed during that period. And also the, um, the PROMISE projects, in addition to others had sort of family.

Effects measured and family based interventions as well. So, there, I mean, they're. Different demonstrations in this long list that have done various things there. Um. I mean, I like the idea of learning from PACE. I would also just mentioned the health professions opportunity grant. Which Laura could say something better she wanted, but at that age project, you know, it's. It's many sites across the country. They're effectively a little mini experiments, a little mini demonstrations happening. Each of those sites. And there are different ways, you can learn from that from the way in which that was done, which could be a model for a future. SSA demonstration potentially, not to mention that that the population of participants there is, you know. Hypothetically potential future entrants or applicants for SSA programs as well.

This is Bob. I'll mention did you want to weigh in on this? Yeah. Yeah, I'll mention just to do that. Come to mind. 1 is the Oregon health insurance experiment that. And the other is the Moving To Opportunity and I think the reason I point to those, too, is that those are two efforts where there is a real emphasis on maximizing the use of the demonstration project, getting the data out to researchers and doing subsequent analysis to maximize the youth. Of the data trying for those 2 experiments. So those are 2 that. In my mind, I'd like to use this kind of models in terms of making. Maximizing the use of the data I'll just say I'm surprised. I think it's interesting that the. No, go ahead. I was just gonna make the observation that the, that kind of. Recommendations that came in response to that question crossed all phases of research, right? From the planning right? And how to be sure that the intervention that you're evaluating is well grounded to begin with.

But then, to your point, Bob, that at the tail end, ensuring that the data ultimately are available to researchers for kind of maxing out on what you can learn after the facts. So, I think that there are definitely opportunities across the potential lifecycle from soup to nuts. If you will to learn from other evaluations and bring those in the Austin, as you pointed out as well, I think that there are. Inklings of these insights already in existing SSA demonstrations. But what else did you want to share here? I was just going to say, I'm surprised that Burt didn't have anything about the unemployment insurance experiments to, to draw on, but. I think there are lessons about intervention design there probably. Um, yes, I think some of the UI experiments have been really interesting and they relate to SSA in some sense, they had to deal with different services and benefits and then restrictions sometimes some, we've been surprised by some of the findings I've worked with people from Abt. They actually on the evaluation of the re, employment eligibility assessment experiment and there I think the biggest problem was that. The design is implemented in the 4 States. We went to was so different. That it was really hard to come up with overall conclusions across the programs and would have been nice to have more systematic designs and more states involved.

Of course, that would have caused more. But that way, we could have learned a lot more than what we did learn, which was a lot, which is that both the penalties and the services both help both help society. And the claimants get back to work. So this highlights some interesting dimensions of testing and intervention and evaluating it where we have. Various population sub populations that were interested in looking at the intervention takes place in multiple places sometimes it's justified in pooling and looking at as a single question and other cases that you highlighted to us and with the, with the HPOG example, that there's so much variation in what the intervention actually is across site that that provides an opportunity to leverage that variation to learn whether there's some particular flavors of the intervention that seem to be more effective than others. And I think we are trying to advance methods to do that. I think that we are at the point where we have no more time left for Q&A, but I want to thank all of the presenters and discussants and Let you know that we are taking a 15 minute break at this point.