INDUCED ENTRY DEMONSTRATION DESIGNS

$1 for $2 Benefit Offset: Exploring the Feasibility of Measuring Induced Entry and Evaluating Potential State-Level Project Designs

September 14, 2000

$1 for $2 Benefit Offset: Exploring the Feasibility of Measuring Induced Entry and Evaluating Potential Project Designs

I. Executive Summary

Section 302 of the Ticket to Work and Work Incentives Improvement Act of 1999 (P.L. 106-170) mandates that SSA conduct a test of the effect that the provision of a $1 for $2 benefit offset would have on the Social Security Disability Insurance program. The legislation specifically requires a thorough evaluation of "the effects, if any, of induced entry and reduced exit from the project." Induced entry is the behavioral response that may result from the implementation of a benefit offset provision in the disability program, that increases the value of the benefit package and encourages persons who would not apply under the existing benefit provisions to apply for, and be awarded, benefits.

From a historical perspective, it is important to recall decisions made in the early 1980's when Section 505 of the 1980 Social Security Amendments mandated a similar test of a benefit offset, without the specific mention of Induced entry. At that time, SSA assembled a large research staff and hired renowned academic researchers as consultants to plan the Work Incentive Experiments (WIE). After careful consideration, but without the thorough investigation undertaken in this research paper, it was determined that it was not feasible to scientifically measure induced entry. Instead, SSA would obtain the best possible measure of the work incentive effects on current beneficiaries and estimate potential program savings. Once savings were known, it would be possible to determine the size of induced entry effect that could be tolerated without eroding all program savings. In the end, the demonstration projects carried out under 505 authority focussed on vocational rehabilitation and other interventions, and the benefit offset was never tested and induced entry never became an issue.

In this paper, we have considered a number of approaches assessing induced entry which fall into one of three categories: experimental methods; quasi-experimental

methods (including state-level comparisons and matched site designs); and analytical approaches (including the analysis of existing data or special evaluative studies). A large-scale experiment was considered to be intractable due to the required sample size and the cost and complexity of tracking these individuals. Four suggested quasi-experimental designs were outlined, and the ability of two of the more promising state-level comparison designs were subjected to a simulation process using recent data on disability awards to determine their ability to successfully measure an induced entry effect. Finally, consideration was given to analytical methods of assessing induced entry, i.e., without direct measurement within a demonstration project.

Key Findings:

A small increase in disability awards associated with induced entry could have a major impact on program costs. The 6 percent increase in awards estimated by OCACT for a $1 for $2 benefit offset for earnings above SGA, and the target effect chosen by the CBO in costing out the legislation, could mean increases in long-run program costs of perhaps $3 billion per year[1]. Because of the very low rate at which persons in the general population enter the SSDI program each year (fewer than 4 per thousand workers) and the fact that even modest induced entry can significantly add to program costs, the study must be designed to detect very small increments in awards (approximately 2 per 10,000 workers). Thus any design that proposes to measure induced entry will require an extremely large sample, and will be very costly (the CBO estimated the cost at $190 million over 10 years). Even small design flaws or unintended study effects could cause the demonstration estimate of induced entry into the program to

[1] The $3 billion long run cost figure is an estimate based upon the lifetime cash benefit and Medicare costs for the 40,000 induced entrants OCACT estimates will be awarded benefits each year. Lifetime cash benefits for a disability beneficiary are estimated to be $100,000 ($700 average monthly benefit X the 138 months OCACT estimates the average beneficiary will stay on the disability rolls). Medicare costs for the average disability beneficiary's entitlement under DI estimated to be about $41,000 (source: OCACT). Thus each year's 40,000 induced entrants could receive a total of $5.6 billion in lifetime benefits. We assumed only partial cash benefits (on average, half the benefit) will be paid to these individuals under the benefit offset provisions, resulting in annual obligations of $3.6 billion. This figure could be higher or lower depending upon the average benefit amount for induced entrants (likely to be higher), the length of time they stay on the rolls (likely to be higher) and how much their DI benefit is reduced by earnings. As a matter of comparison, OCACT estimated that, in the short run, over the first 5 years, induced entrants would add $4 billion to SSDI program costs (excluding Medicare). The short run costs rise significantly over time.

be significantly different than the effect that would occur in a nationally implemented program.

All demonstration designs were found to suffer from the potential for unintended study effects that could result in inaccurate estimates of induced entry. First, getting adequate information to those who are included in the demonstration so that they could potentially take advantage of the offer, but avoiding providing the wrong information to those not included in the demonstration was a major concern. This was particularly a concern for demonstrations employing localities, such as counties or field office service areas. However, even in State-level demonstration projects there was concern that the media market area would extend beyond state borders, failing to confine information to the population in the project. Second, migration from a location not offering the benefit offset to a location in the project was also a concern. Third, the fact that there may be a "stock" of persons who would find the benefit offset more attractive than the current program could result in initial levels of induced entry that are higher than the long-term rate of induced entry based upon the "flow" of persons who would be induced to enter. A fairly long demonstration intake period may be required to disentangle the short-run stock induced entry effect from the longer-run flow induced entry effect. In conclusion, there are study effects that can be expected to under-estimate induced entry, and others that will over-estimate induced entry. There is no way, a priori, or post-priori to determine the relative size of each effect and whether the net impact of the study effects produced an under- or over-estimate of induced entry.

A number of statistical issues were also identified. First, it became clear that an experimental design based upon individual assignment and micro-level analysis was not feasible due to the size of the project (several million persons each in the control and treatment groups), and the study would have to rely on a quasi-experimental design utilizing aggregate data, such as award rates. Two potential quasi-experimental designs, one a matched-state design and the other a 5-state design, were tested as to their ability to measure an induced entry effect. Simulations were conducted utilizing actual state-level awards rates in a differences-and-differences (pre-post and comparison-state) analysis approach, with simulated induced entry effects imparted in the data. These simulations show

that the existing variation in awards rates across states is too large to permit accurate estimates of induced entry effects, even in the best case scenario where controlled, fixed effects are embedded into the data. Thus even a large scale costly ($200+ million) demonstration project can not be reasonably expected to provide policymakers with accurate information about induced entry to assist in a decision about the cost-effectiveness of a benefit offset.

The Proposed Approach:

In lieu of designing a demonstration to "measure" induced entry, we propose first examining the behavior of current beneficiaries under the $1-for-$2 benefit offset demonstration projects. The results from the proposed demonstrations will provide estimates of the savings generated by the return to work of current beneficiaries. These savings estimates will provide a baseline measure for the assessment of the impact of induced entry. Specifically, we can determine whether the savings from beneficiaries' return to work are sufficient to cover the potential costs of our analytical measure of induced entry described below. We can then assess the likelihood that the implementation of a benefit offset will result in program savings.

We will use several approaches to assess the actual potential for induced entry. Employing multiple analytical approaches will help assess the sensitivity of the estimate to the approach and inherent assumptions and help us work towards a reasonable assessment of induced entry. These alternative approaches would include analysis of existing data sources (such as the National Study of Health and Activity (NSHA)), actuarial estimates, a special survey, and small-scale experiments analyzing individual decision making behavior.

The NSHA will contain critical information needed to assess induced entry. The NSHA is a nationally representative study of the prevalence of disability in the general population, with a specific goal of identifying and enumerating those who would qualify for SSDI and SSI benefits if they chose to apply. The existing study design of NSHA will make it possible to estimate the number of people who are eligible for SSDI. This estimate may be used as an upper bound of the number of induced entrants because it is based upon the strong assumption that

everyone who is eligible for the program would enter the program under the $1-for-$2. It is unlikely that everyone who is eligible for the program would apply under the $1-for-$2. The more difficult task will be to determine the number of persons who are eligible for benefits and who would actually be induced to apply for benefits by the $1-for-$2. This estimate will be made by examining how beneficiaries make the decision to enter the program.

The decision to enter the program will depend upon how a person values non-beneficiary status compared to how they value entry into the DI program. If their value of non-beneficiary status is greater than their value of entry into the DI program they will choose to delay entry into the program. If the value of the DI program is greater than the value of non-beneficiary status, they will apply. The value of non-beneficiary status will be estimated by their non-beneficiary income and other benefits that they receive as a non-beneficiary (e.g., pensions, health insurance, union status, etc.). The value of entry into the program will require much more work to construct because it will depend upon:

- their income as a beneficiary,
- their value of leisure time,
- the cost of foregoing Substantial Gainful activity during the five month waiting period,
- the cost of potentially foregoing SGA for a longer period of time during the decisions process,
- the risk that they might be denied benefits at some point during the disability determination process,
- the value of other benefits received as a beneficiary (e.g., Medicare benefits after a 2-year waiting period, other benefits that they might receive from work while in beneficiary status), and
- Other factors associated with the value of DI benefits.

The NSHA survey data, administrative data, and the NSHA estimate of the number of non-beneficiaries who are eligible for the program will be used to develop these estimates.

Several different techniques will be used to obtain reasonable estimates of these values. These alternative techniques use different assumptions to place a value on non-beneficiary status and beneficiary status. One method

will use structural econometric models that use NSHA data to obtain estimates these values.

Another method, based upon a different set of assumptions, would utilize interviews with potential beneficiaries. These interviews involve obtaining the person's value of beneficiary and non-beneficiary status through a set of carefully designed questions. This information will supplement NSHA data to form another estimate of induced entry.

A third method would be based on assumptions used by the Office of the Chief Actuary (OCACT). OCACT constructed estimates of induced entry prior to the enactment of the Ticket to Work and Work Incentives Improvement Act legislation. The new information provided by NSHA, as well as information from the proposed $1-for-$2 demonstration, might provide OCACT with new information that will allow them to update their estimate of induced entry.

Attacking the problem from different angles with different sets of assumptions will, at best, provide a robust estimate of induced entry. At the very least, it will provide different sets of estimates of induced entry based on different assumptions. In this case, the "best estimate" will be determined based upon a careful assessment of the assumptions underlying each approach.

Conclusion:

Based on the analysis presented in this paper, it is concluded that it is not worthwhile to undertake demonstrations that purport to measure induced entry, as the resultant measure will be of unknown accuracy. In the paper, we first outline the statistical issues and study effects involved in a demonstration project designed to measure induced entry. We conclude from the analysis that the sample size required for statistically reliable results of induced entry would be extremely large. For example, well over 1 million people would be required for a treatment group, and another million for the control group, to statistically identify the induced entry effect that both CBO and OCACT stated had very important implications for the program. A randomized experiment with over 1 million in a treatment group, and the same number in a control group, is much too large to be practical. Even if the resources were available for such a design, additional

resources would be required to minimize the impact of biases that we show might exist from the potential study effects inherent in the design.

We then perform simulations, based upon different state-level research designs, using actual SSA awards data that would be used for the evaluation. In this design, an entire state, or set of states, would be eligible for the $1-for-$2 benefit. Induced entry would be measured by comparing the awards rate in one state (or set of states) that receives the offset to the awards rate in comparison state (or set of states) that does not receive the $1-for-$2 benefit. The key to this type of analysis is that the comparison state(s) has to be able to isolate the impact of the $1-for-$2 offset from the many other factors that affect the awards rates. In these simulations, we place an induced entry effect in the data and perform the analysis that would be done based upon the design. Therefore, we were able to create the type of situation that is likely to occur if one of these demonstration designs were actually carried out in order to measure induced entry. We found that the simulated evaluation could not consistently obtain reliable measures of the induced entry effect we created in the data. Again, it is very important to add that these simulations are based on a best case scenario; that is, they have none of the problems associated with the study effects discussed in the following document. The details of these simulations are shown in Appendices D and E. We conclude from this analysis that a state-level research design would provide unreliable results due to the problem of isolating impact of the $1-for-$2 benefit offset from other factors within a state that affect the awards rate. The resource required to obtain this unreliable result from a state-level design will be very large. Therefore, we do not recommend that SSA pursue a state-level research design.

Finally, we propose alternative analytical approaches that will use SSA's NSHA data to examine the potential for induced entry. Using the different proposed approaches, with different assumptions, we can obtain a set of estimates of induced entry. These estimates will be examined and a "best estimate" will be determined. The "best estimate" is likely to be more reasonable than an estimate generated from suggested induced entry demonstration projects that suffer from both statistical shortcomings and study effects. It will also require fewer

resources and be a considerably less expensive approach to achieving a potentially better estimate of induced entry.

II. The Legislative Mandate

The Ticket to Work and Work Incentives Improvement Act:

Section 302 of the Ticket to Work and Work Incentives Improvement Act (TWWIIA) of 1999 (P.L. 106-170) requires the Social Security Administration (SSA) to conduct a set of demonstration projects that reduce the financial disincentives to work and earn income in the Social Security Disability Insurance (DI) program. It states that:

> "The Commissioner of Social Security shall conduct demonstration projects for the purpose of evaluating, through data collection, a program for title II beneficiaries under which benefits payable . . . are reduced by $1 for each $2 of the beneficiary's earnings that is above a level to be determined by the Commissioner."

The demonstration projects "are to be conducted at a number of localities that the Commissioner shall determine is sufficient to adequately evaluate the appropriateness of national implementation of such a program."

Section 302 lists six objectives for the demonstration projects:

- Measuring the extent of induced entry and reduced exit from the Disability Insurance program,
- Determining the relationship between the demonstration projects and the Ticket to Work and Self-Sufficiency Program,
- Measuring the savings that may accrue to Social Security and other federal programs,
- Determining the annual cost of the demonstration project and the annual cost that would have been incurred in the absence of the project,
- Identifying the determinants of return to work, including characteristics of the beneficiaries who participate in the project, and

- Assessing the employment outcomes, including wages, occupations, benefits and hours worked, of beneficiaries who return to work as a result of the project.

In designing the projects, SSA is required to take into account the advice of the 12-member Ticket to Work and Work Incentives Advisory Panel. Members of the panel are to have experience or expert knowledge as a recipient, provider, employer, or employee in the fields of or related to employment services, vocational rehabilitation services, and other support services.

The legislation requires that:

> "The demonstration projects developed under subsection (a) shall be of sufficient duration, shall be of sufficient scope, and shall be carried out on a wide enough scale to permit a thorough evaluation of the project to determine—
> (A) the effects, if any, of induced entry into the project and reduced exit from the project"

It is this requirement that served as the genesis of this paper. While it was apparent that demonstration projects could be developed to assess reduced exit from the program (and many of the other objectives listed), it was not obvious how SSA might develop a project to assess induced entry. This paper is intended to objectively assess the problem, past research on the subject, literature pertaining to induced entry, and the potential to develop a project to measure induced entry.

The Social Security Disability Amendments of 1980:

TWWIIA was not the first time Congress expressed interest in assessing the impact of a benefit offset on the DI program. Section 505 of the Social Security Disability Amendments of 1980 (P.L. 96-265) contained demonstration authority closely paralleling the TWWIIA authority. SSA conducted a number of demonstration projects under that authority, but did not get beyond planning stages for a test of a benefit offset. Section 505 did not have an express requirement to assess the effect that a benefit offset would have on induced entry. Demonstration plans, as discussed below, focussed predominately on the impact of a benefit offset on work behavior of current beneficiaries.

III. Induced Entry: The Issue

What is Induced Entry:

Induced entry is the behavioral response that occurs when a change in policy, such as the implementation of a benefit offset provision in the disability program, encourages persons who would not apply under the existing benefit provisions to apply for, and be awarded, benefits. In general, certain policy changes may increase the attractiveness of the program (i.e., increase the value of the benefits package) and provide greater incentives for non-participants to become participants. In the present context, the addition of a benefit offset provision to the SSDI program under which benefits (albeit reduced) can be collected despite substantial earnings changes the nature of the program from a total disability program to a partial disability program, where benefits are supplements to earnings. This clearly increases the attractiveness of the program for some persons, and non-beneficiaries may choose to become beneficiaries. The size of the induced entry effect depends on 2 things: (1) the number of individuals that are not currently participating in the SSDI program who qualify for disability benefits and (2) the generosity of the benefit offset and Medicare provisions. Non-beneficiaries who qualify may have chosen not to apply for any number of reasons, including having a job with earnings exceeding SGA, that make the current structure of disability benefits or program requirements unattractive.

Induced entry, other things equal, can only lead to increased program costs. The program cost of induced entry effect will be greater for more generous benefit provisions than for less generous provisions. For example, a larger induced entry effect can be expected from benefit offset has a relatively high earnings disregard or a low rate of offset. Unfortunately, the benefit offset provisions that offer greater incentives for return to work by current beneficiaries, also have a greater potential for costs associated with induced entry. In other words, the more we offer to get current beneficiaries to work, the more persons will find the program attractive and seek to become beneficiaries.

The Impact of Induced Entry:

The addition of a benefit offset to the SSDI program is intended to increase incentives for beneficiaries to seek work and to result, through reduced benefit payments, in

program savings. As noted above, such a provision will also make the program more attractive to non-beneficiaries, and the resulting induced entry will raise program costs, reducing potential savings or perhaps increasing program costs overall. It is a delicate balancing act to increase incentives for current beneficiaries to work, while minimizing the potential for induced entry, in order to maximize program savings.

The Office of the Chief Actuary (OCACT) estimates that a $1 for $2 benefit offset applying to earnings above SGA level will result in 40,000 new awards (induced entrants) per year. That increase (approximately 6%) in disability awards could increase long-run program costs (cash benefits and Medicare) by perhaps $3 billion per year[2]. OCACT also estimates that a $1 for $2 benefit offset applying to earnings above an $85 disregard, as in the SSI program, will result in induced entry about half that magnitude. When comparing the costs of induced entry to the savings accruing to additional work by current beneficiaries, OCACT estimates that both provisions will result in no savings, but rather higher overall program costs.

Clearly induced entry plays a large role in the potential success of a benefit offset provision and must be taken into consideration. The question is how to design a project that meets the legislative mandate to measure induced entry.

Historical Perspective- The Work Incentive Experiment:

As a historical perspective, it is interesting to consider the planning and design of the Work Incentive Experiment (WIE) that was developed under Section 505 of the Social Security Disability Amendments of 1980, but was never carried out. Section 505 had no specific mandate to evaluate induced entry. SSA staff and a panel of (8) distinguished academic researchers[3] hired as an advisory board carefully considered all WIE objectives, evaluation requirements and design issues and developed a plan to test, among other provisions, a $1 for $2 benefit offset for earnings above SGA level. During the deliberations,

[2] See footnote 1.

[3] The consultants on the advisory board included: Orley Ashenfelter (Princeton University), Robert Baruch (Northwestern University), Joseph Sedransk (SUNY Albany), Burton Singer (Columbia University), Donald Patrick (St. Thomas's Medical School, London), Glen Cain (University of Wisconsin), Lee Rainwater (Harvard University, Robert Groves (University of Michigan)

the issue of induced entry arose. While the importance of induced entry and the potential impact on costs was acknowledged, it was decided that there was no practical way to design, and conduct, a project that could measure induced entry. Instead, SSA should focus first on the behavior of current beneficiaries to determine the potential savings from return to work. If significant savings were found, then induced entry could be evaluated using alternative analytical techniques, within the context of these savings.

There is little documentation from the 1980's project identifying the specific considerations that resulted in dismissing a project design intended to measure induced entry. Certainly there was no formal analysis as provided in this paper. However, the February 1981 WIE research protocol states:

> "Finally, there is no mechanism within the current experimental design to assess the effect on trust fund costs of the possible increase in applications for disability benefits due to the attractiveness of the programs embodied in the experimental alternatives. This issue will be addressed in separate analyses or experiments."

IV. Parameters of a Demonstration Design and Evaluation

The parameters of a specific project need to be defined in order to properly specify an evaluation design. The key elements of such a design are described in Hollister and Hill (1995). This section defines the parameters of a design that measures induced entry from a benefit offset provision.

What is the Research Question? The legislation requires a research project that is of sufficient duration, sufficient scope, and that is carried out on a wide enough scale to answer the following research question put forth by the legislation,

"What are the effects, if any, of Induced Entry into the project and reduced exit from the project?"

The legislation also requires an estimate of induced entry that would result from permanent implementation of such a program.

What is the Target Population? The target population includes all persons between the ages of 18 and 64 in the population that might apply and be awarded benefits at some point in time during the demonstration period. It is possible for anyone to experience a health condition that prevents Substantial Gainful Activity at any time. Therefore, the target population includes all persons between the ages of 18 and 64 who are not currently collecting Title II benefits.

Establishing benefit offset provisions to be tested that will be representative of a national program. If participants are assigned to a treatment group that does not have provisions that are consistent with those that would be in place in a national program, the demonstration results will not replicate the effect of national implementation. If demonstration participants do not receive sufficient information with respect to the "new program" or do not understand the benefit offset provisions that are being offered, they cannot be expected to respond to the provisions. It is extremely important that each demonstration participant has a clear understanding of the benefit provisions, and that the offer mirrors the proposed "new program". Demonstration provisions that differ from the proposed program, and a lack of adequate information or understanding, may lead to results that are not representative of the effect that the benefit offset provisions would have on induced entry. (These issues are discussed more fully in the latter section on study effects.)

Constructing the Counterfactual. In order to measure the impact of a program on a specific outcome, a measure of the outcome in the absence of the program, referred to as a counterfactual, is required. In a randomized experimental design, the counterfactual is measured using a control group. In non-experimental designs, such as a matched site design, the counterfactual is established by providing a basis for comparison, generally a control or comparison group. Hollister and Hill (1995) describe the issues associated with selecting a comparison group.

In this paper, persons who are subjected to the benefit offset provisions are referred to as the treatment group. Another group of persons who are subjected to the existing program is used to form the counterfactual. This group is referred to as the control group under a randomized

experiment and a comparison group in other non-experimental designs.

What is the Outcome of Interest? The research question is whether the treatment (benefit offset) causes some individuals to reassess their desire to apply for disability benefits and whether the individuals who are induced to apply are actually awarded benefits. On an individual level, the question is whether the probabilities of application and award increase and, for individuals who are awarded benefits, what is their labor supply behavior and benefit level. Overall, estimates of the number of additional awards, and the work behavior and benefit costs, are needed to determine the increased program size and cost associated with the new program provisions. Therefore, making an estimate of the size and cost of induced entry requires demonstration estimates of the following outcomes:

1. The number of persons who are offered the new program (i.e., are in the benefit offset group) who apply for the program and are awarded benefits.

2. The number of persons who apply under the existing program. (measured using the control/comparison group)

3. The length of time that persons who are offered the new program (i.e., are in the benefit offset group collect benefits as well as the amount of those benefits.

4. The length of time that persons collect benefit under the existing program as well as the amount of those benefits. (measured using the control/comparison group)

The awards rate (defined as the number of persons awarded benefits in the group divided by the total number of persons in the group) is the outcome used to measure the number of persons who apply and are awarded benefits in the treatment and control group for a given year. However, because induced entrants may be quite different in terms of labor supply than those who enter the rolls in the absence of the new treatment, it will be necessary to assess the overall labor supply and benefit costs of the two groups. Thus estimates will be required of the average number of years on the benefit rolls and the average benefits paid, over time, for both the control and treatment group.

V. STATISTICAL ISSUES

How large does induced entry have to be to have an important impact on the program? The effect that has been most frequently cited as an estimate of induced entry is 40,000 new awards per year (6% percent increase in the awards rate) for the $1 for $2 benefit offset of earnings above SGA level. Both the CBO and the OCACT used this level in their cost estimates. Induced entry from the $1 for $2 benefit offset for earnings above $85 was estimated to be half as large, or about a 3% increase in the awards rate. These cost estimates were viewed by Congress as too large or too uncertain to permit legislation establishing a benefit offset for the SSDI program and it was mandated that SSA test the offset. Congress did not specify what level of induced entry, and program cost, might be tolerated within the current budgetary climate and it was difficult to determine what effect size must be detected. Lacking more specific information about the effect size of interest, we assumed that we would have to detect differences somewhat smaller than those cited above. We based (somewhat arbitrarily) the sample size requirements on power analyses that would permit the detection of a 2 percent increase and a 5 percent increase in the awards rate. (These correspond to an annual long-term increase in program costs of roughly $1 billion and $2.6 billion, respectively.)

Determining the sample needed to detect a 2 percent (or 5 percent) increase in the awards rate. The awards rate in the target population is very small—on the order of around .003 (3 per 1000 persons) for a given year. If the project is able to identify insured workers as a target population, the awards rate is still very small between 0.0045 to 0.0055 (4.5 per 1000 or 5.5 per 1000) for a given year. A 2 percent increase in the allowance rate in either case is an extremely small change (on the order of .00006 to .00011) to find, but this small effect will have extremely important implications for program costs. Detecting such a small change requires a very large sample.

A power analysis was used to estimate the sample size necessary to detect a 2 percent change in the awards rate as well as a 5 percent change in the awards rate. Power, in this context, refers to the probability of detecting a specified effect when the effect actually exists. In our

calculations, we used two power values, 0.80 and 0.90, because we wanted an 80/90 percent chance of detecting the effect if it actually exists. A larger sample is required to increase the power of the evaluation (holding all other assumptions constant).

To perform the power analysis, we used the parameters outlined in this document and the assumptions described in Appendix Tables A and Appendix Table B. A good description on power calculations is in Cochran (1983 p.50-73) and Cohen (1977).

To find the small, but costly, change in awards, an extremely large sample is required. Our estimates in Appendix Table A show that for a randomized experiment, a sample on the order of 9 to 12 million would be required for both the treatment and control group to find the effect. These power calculations are based upon the use of a simple random sample from the population; other sampling techniques (such as a clustered sample) may require larger sample sizes. In any case, sample sizes of this magnitude make a randomized experimental design unmanageable for the purposes of testing induced entry. While the sample can be reduced under different assumptions specified in Appendix Table A, the required sample sizes are still extremely large.

Sample Design Effects. The sample sizes above are based upon the assumption of a simple random sample. However, it is likely that that such a demonstration of induced entry would require a clustered design (based upon geographic locations such as states or counties) to facilitate effective control of, and data collection for, the demonstration project. Limiting the randomized sample to certain geographic areas, know as clustering, will result in design effects that may change the efficiency of the sample design relative to that obtained under a simple random sample design. In many cases, clustered designs are less efficient than a simple random design and require larger sample sizes to compensate. Thus, to achieve the same statistical precision the number of persons offered the treatment and tracked for project is likely to increase. And in order to avoid problems with quasi-experimental designs, individuals within each geographic area would need to be randomly assigned to the treatment and control group, again raising issues of feasibility and manageability. It is unknown at this time what the design

effect might be and SSA intends to have its consultants and design contractor address this issue, however it is not unlikely that the sample size could increase by 50 to 100 percent.

Measurement error – Bias from Quasi-experimental design. In a quasi-experimental design, individuals are not randomly assigned to treatment and control within the overall population or clusters, but instead a comparison group is achieved by matching and randomly assigning (for example) comparable areas in a clustered design to the treatment or control. While this improves the ability to control the project, especially information given to members of each group, there is a serious issue with respect to the validity of the results. Variation between the locations selected, and individuals residing in the location, in factors that influence demonstration outcomes may produce effects in addition to the effect produced by the treatment, and contribute to the overall measured difference in outcomes between the two groups. This confounding effect, or bias, can cause project results to not only fail to detect the true effect, but it may also suggest an effect that is of a different magnitude or even sign from the underlying impact. Thus, such a design runs the risk of not only failing to detect the true effect, it may suggest a favorable impact when the true effect is harmful. There is no way to determine absolutely whether the true effect was detected.

Robinson G. Hollister and Jennifer Hill, in their paper "Problems in the Evaluation of Community-Wide Initiatives", warned:

> *It is important to stress, once again, that the vulnerability to bias in estimation of the impacts of interventions should not be taken lightly. First, the few existing studies of the problem show that the magnitude of errors in inference can be quite substantial even when the most sophisticated methods are used.* ***Second, the bias can be in either direction: we may not only be led to conclude that an intervention has had what we consider to be positive impacts when in fact it had none, we may also find ourselves confronted with impact estimates which indicate, due to bias, that the intervention was actually harmful; we may be misled either to promote policies which in fact use up resources and provide few benefits or we may be led to discard types of interventions as unsuccessful which actually have underlying merit.*** *Once such biased quantitative findings are in the public domain, it is very hard to get them dismissed, to prevent them from influencing policy decisions, even when we have strong intuition that they are biased." (p. 27)*

Measurement Error- Level of analysis. Although the behavior we wish to measure occurs among individuals, a project of this size cannot possibly track all individuals offered the treatment in order to measure the change in behavior. Instead we must rely on upon aggregate measures, such as the award rate, to assess the impact of the treatment in terms of induced entry. This leads to two sources of measurement error. The first is aggregation bias, where summing up individual behavior into one aggregate measure may mask the underlying micro-relationships and result in biased measures. This is particularly important when doing a quasi-experimental design and differences in conditions and participants across sites may contribute to the outcome, but due to the aggregation it is not possible to adequately adjust for these differences. The second source of measurement error occurs from the fact that there is inadequate information about the aggregate group. For example, in any county or state-based evaluation, which utilizes award rates as an measure of induced entry, there is no systematic and accurate data on the number of persons (size of the population) at risk, which forms the denominator for the award rate, and how that number is changing over time due to migration, death, etc. Even at the State level, the decennial census is the best source of information and those counts suffer from measurement error and are not updated over time. The resulting inaccuracy in the denominator could produce errors in the award rate that could significantly affect the comparison of award rates across locations, particularly where the variation is systematic. Although SSA can directly count the awards (in the numerator from administrative data, this measure will not reflect study induced changes, such as migration to the locale to take advantage of the treatment (see study effects below).

Existing variation. Period to period variation exists within all locales, as well as across locales. Some variation may be the effect of certain events or underlying factors in the population (explained) or simply random variation (unexplained). Using micro-level data it is sometimes possible to adjust for explained variation. Using aggregate data it is impossible to explore the underlying relationships and account for some of the existing variation. The existing variation in aggregate awards rates among States over time has been examined and the indications are that existing variation may be larger

than the effect size of interest. This could mean that any induced entry demonstration will fail simply because the existing variation (that not due to the treatment) will overwhelm the effect of the demonstration and result in no detectable change. Based upon State-level data, we believe that this could occur even if the induced entry effect is larger than the effect deemed important to the program from a cost perspective (see Appendix C). Variation in smaller geographic areas (e.g., counties) is even larger, and is more likely to mask any induced entry effect.

Model-based approaches to adjust for differences in the factors that influence the award rates, whether done on a micro- or macro-level may also be problematic. The same variables that are used to adjust for variability across areas may also be factors that play a role in determining induced entry. This multicollinearity may result in overstating or understating the induced entry effect. A strong micro-level "difference-in-differences" approach tends to minimize this problem, however differential trends across treatment and comparison sites could result in significant biases in the measure of the treatment effect. This potential bias is likely to be a greater problem in the aggregate data approach, as the underlying relationships are not accounted for and only net impacts are measured.

VI. Study Effects

For the induced entry demonstrations, study effects refer to errors in the measurement of induced entry due to a design that:

1. Does not adequately measure behavior that would occur in the absence of the project (Imperfect Comparison group described in Hollister and Hill (1995)), and/or

2. Does not mimic the conditions that would occur in a nationally implemented program.

The impact of study effects may not be directly measure-able. It is important to identify the potential study effects associated with a particular design prior to the implementation of a demonstration. Identification of potential study effects allows one to choose a design that minimizes the impact of these effects and allows one to

assess the implications of the study effects on measures of induced entry.

Inducing the behavior

Information and understanding- Individuals in the induced entry treatment group must be provided with complete and accurate information, and understand the new program in order to be able to react to the availability of the treatment. Without information, the study effect will be to bias the induced entry effect lower than what might exist in an ongoing program. On the other hand, providing potentially eligible persons with information that they might not otherwise receive may, in itself, cause some persons to apply imparting an upward bias in the estimate. Hence, similar information on the provisions of the current program must be provided to the control group to minimize the bias[4]. If an estimate of the induced entry to the program from this informational component is desired, a blind control group, i.e., not receiving any additional information, must be created.

Past research[5] performed by SSA shows that the decision to apply for benefits is not always an individual decision, but is often based upon the advice of others such as doctors, government agencies (e.g., welfare department), relatives and other "advisors." It will be difficult to provide information about the availability of the treatment or "new program" to applicant advisors to facilitate the dissemination of information to potential applicants.

Time to respond- Little is known as to how and when disabled individuals decide to seek benefits, although there is some evidence that applications rise during periods of increasing unemployment. It will be necessary to provide sufficient time for those eligible for the treatment to be "induced" to apply. Insufficient time will bias the measure of induced entry downward. After allowing sufficient time for induced entry to occur, it will be necessary to provide a reasonable observation period to determine the differences in labor supply and benefit costs

[4] A third study group, a blind control group, may be required to assess the impact that information about the program and its provisions had on application behavior. The method of informing both groups to isolate the induced entry effect relies upon an assumption of additivity, i.e., that the effect of information is the same for both groups. If there is an interactive effect between the treatment and information, there may still be a bias in the measurement of induced entry.

[5] See, for example, the results of SSA's 2-day Survey of Applicants.

while on the benefit rolls. (see also stock vs. flow in the measurement study effects below)

Trust- Providing information to potential eligible so they understand the provisions of the new program is only a part of what needs to be provided. This demonstration project is not the same as a nationally implemented program and may be viewed those involved as "subject to change." Individuals receiving the offer must be convinced that the offer that has been presented will be honored on an ongoing basis. Without implicit trust that SSA will adhere to the offer and not renege, the individual may not accept the risk of stopping work in order to apply for benefits, thus understating the induced entry that would occur in an ongoing program.

Time limitation on access to treatment- Any induced entry demonstration will necessarily be time-limited in order to facilitate evaluation and to limit exposure to administrative and program costs. This time limitation may induce more applications than would otherwise occur due to the "act now, or miss the opportunity of a lifetime" effect. This study effect will tend to overstate the true induced entry effect, and will be even more exaggerated when combined with the stock/flow measurement problem discussed below.

Measuring the behavior-

Stock vs. flow- The cost to the DI trust fund of offering a program with a benefit offset will include the costs associated with increased entry and decreased exit. It is likely that there currently exist a number of persons who would benefit from the entering a program with a benefit offset, and a number of beneficiaries on the rolls who would work if given a benefit offset. These individuals represent the stock of individuals who would currently find these provisions advantageous. Overtime there will be a number of persons who become disabled and might find these provisions more advantageous than the current program. This is the flow. The costs of induced entry can be expected to change over time as induced entry from the existing stock gives way to an ongoing flow of persons who reach the point where the offset would be advantageous to them (either because their disability gets worse or their economic circumstances change).

The presence of the stock of persons who would find the new program attractive will lead to a short run level of induced entry that will likely exceed the long run (equilibrium) level of induced entry. Thus the induced entry measured early in the project may overstate the long run level of induced entry, requiring a significant time period for the project to permit absorption of the stock and to be able to observe the longer term flow. Clearly the short run induced enter will raise program costs in the short run, but to make longer term estimates of program costs the equilibrium level of induced entry must be known. The length of time required to obtain accurate measurements of the long run level of induced entry is unknown; we expect to see induced entry peak, decline, and then level off. It is that final level of induced entry that will be the best measure of the longer run, and only through observation and measurement over time will it be apparent that the project is reaching the equilibrium rate of entry.

Migration- Measurement problems relating to aggregate measures of award rates were discussed earlier. Clearly, utilizing a quasi-experimental design utilizing populations in set geographic areas means the population provided the treatment can change through migration from area to area. Some of this migration is normal (in a two year period approximately 16% of the US population moved, 10% within the same county, 3% to a different county in the same state and 3% to a different state). However, offering the incentives of a benefit offset in some locales may induce some persons to move into those locales where the benefit is offered so that they may benefit from the provision. Unless we can completely control the flow of information about the project and the specific individuals who receive the offer of the treatment, it will be extremely difficult, if not impossible, to separate the "normal" level of migration from the induced migration. In the aggregate analysis of allowance rates, even employing differences and differences (cross-site, pre-post measures) approaches, the increased applications from individuals outside the area will overstate induced entry. This problem is best summarized by Hollister and Hill (1995):

"In and out migration of individuals are a constant feature in communities. …. In migration could be due to the increased attraction of services or it could be a natural process which will weaken the homogeneity of community values and experiences. Out migration means the loss of

some persons subject to the treatment. If one looks only at the stayers in the community there is a selection bias arising from both migration processes. One cannot be sure the program treatment itself influenced the extent and character of in and out migration."

Ethical Issues

Such a test would clearly require review for Human Subjects Protection and, even if approved, may still raise ethical issues. Involving non-beneficiaries in the study is quite different from studying the effects on beneficiaries. Rather than encouraging individuals to increase work effort and reduce reliance upon income support program, this aspect of the study has the potential to reduce work effort and increase reliance upon such programs. Offering this treatment to individuals who are not currently on the rolls is intended to measure the number who would choose to forsake (or reduce) work for the opportunity to apply for benefits with the potential for receiving a partial disability benefit while working. Some individuals may leave jobs in order to facilitate an application, but not qualify for benefits. Such decisions, and the resulting disruptions, could result in long term consequences to the worker and family. Information provided by any of a number of methods (media, word of mouth, etc.) to those not included in the test (i.e., in neighboring locales) could lead to misinformed decisionmaking, and ethical issues that might be resolved by the legal system with potentially huge costs to the trust funds.

Potential for Administrative and Program Costs

In order to measure induced entry (or to attempt to do so) SSA will be required to undertake a test that creates the behavior that we intend to measure. This means that SSA must permit an undefined group of unknown size with the opportunity to apply for benefits and, if they qualify for benefits, to receive benefits under the test provisions until they recover, attain age 65, or die. There may be very high costs, both program and administrative, associated with such a test. Whether the individuals qualify for benefits or not, SSA must process disability claims which represent an administrative cost. If the individual qualifies for benefits, SSA will pay lifetime, albeit reduced, cash benefits and offer Medicare to an individual who would not have otherwise applied, at least

at that point in time. The average cash value of lifetime benefits is nearly $100,000, not including Medicare, and could result in a very costly project if large numbers of individuals are induced to apply. Furthermore, program costs for other income support programs may rise as individuals seek unemployment insurance, welfare, and other support during the waiting period, and continue on these programs because they are not awarded disability benefits.

VII. Possible Designs to Measure Induced Entry

Below are several designs that could be considered as a method to measure induced entry. For each design, there is specific discussion of the various statistical issues and design effects that could affect the validity of the measure of induced entry.

Method 1. Random Sample from the Target Population

Design. This research design uses persons between the ages of 18 and 64 as the population that might apply and be awarded benefits at some point in time during the demonstration. The design selects a sample from this population and randomly assigns them to either the benefit offset provisions (treatment group) or to the existing benefit provisions (control group). The impact of induced entry would be measured on an annual basis as difference between the proportion of persons who apply for benefits in the treatment group and the proportion of persons who apply for benefits in the control group. The difference in proportions would be used to produce an estimate of the number of awards made in the population on an annual basis due to the benefit offset provisions.

Tracking Participants. If it were possible to identify and inform both demonstration participants and a control group, the evaluation would require tracking all individuals in both groups to assess applications for the program and awards. While the sample to be tracked is large, it is feasible to follow application and awards behavior for participants in the demonstration.

Required Sample Size. Appendix A shows the sample size required under various assumptions to detect at least a 2 percent increase in the number of awards and a 5 percent increase in the number of awards. Attempting to measure an

effect size consistent with that identified by CBO and OCACT (2 percent increase in awards or $1 billion annually) and lowest acceptable statistical power (.80), **a sample of 9 million persons for each group is required.** Even under a design with the smallest level of statistical power (.80) and a large detectable effect (5% increase or about $2.6 billion annually), **a sample of 1.44 million persons for each group is required.** Increasing the power to .90 requires a larger sample.

Informing Participants. Methods to provide information about the benefit offset that might be used in a nationally implemented program (e.g., employing the mass media) cannot be used in this demonstration design because the information cannot be confined to persons assigned to the treatment group. Information on the program would need to be provided on an individual-by-individual basis, which would be difficult and resource-intensive undertaking under the demonstration, and would not mimic the information that would be received under a nationally implemented program.

The method of informing participants would involve providing each individual with information, either through face-to-face interviews, detailed telephone explanations, or by mail. In order to provide sufficient information on the benefit offset provisions for the treatment group and the existing provisions for the control group, with an opportunity for questions, we would propose a face-to-face explanation. Repeated interviews might be required in order to remind persons in the demonstration of the provisions. Even in the smallest demonstration, with 1.44 million in each group, a detailed face-to-face explanation so that each participant understands benefit provisions would be too expensive to perform. Detailed personal telephone calls that carefully explain the benefit offset provisions to each participant would not be able to cover the entire sample, as some have no phones, and others unlisted numbers. Phone contacts would also be very expensive.

Less expensive methods such as mailings and automated telephone calls will likely be lost, ignored or mis-understood by many participants. As a result, sufficient information may not be provided to participants. Without sufficient information, participants cannot respond to the benefit offset provisions. The results under such an approach could severely understate the impact of induced

entry if the program were to be administered on a national level.

In no case would information about the "new program" available to this individual be received by third party advisors: persons, such as doctors, government agencies, relatives, and others who assist the individual in making the decision to apply for disability benefits. This would likely result in induced entry being understated.

Study Effects. The following study effects are possible in this research design:

- The provision of information to participants cannot replicate that which would occur in nationally implemented program. This might lead to a different induced entry behavior compared to what would occur in a nationally implemented program.

Conclusion. This design would not suffer from the statistical issues of a quasi-experimental design. However, the sample required to detect an induced entry effect is much too large to make a randomized experiment feasible and individual data collection would be cost-prohibitive. Dissemination of information to those eligible would be extremely costly and may not reach all actors in the application decision.

Method 2. Matched Counties (Matched Geographic Locations)

Research Design. The matched county design first selects a sample of paired counties or similar geographic entities. The counties are paired so that they are as similar to each other as possible in terms of the factors that affect the outcome that is being studied. For example, counties may be matched based on past awards rates and economic factors in order to measure the impact of the benefit offset provisions on the awards rate. After counties are matched to form pairs, one of the counties in each pair is randomly selected as a benefit offset provisions site (the treatment) and other county receives the provisions in the existing program (the control).

The impact of the benefit offset provisions is measured by first forming the difference in awards rates between the treatment group and the control group in each matched pair.

If the two groups have different awards rates prior to the introduction of the benefit offset provisions, then the change in awards rates before the benefit offset provisions are implemented and after the benefit offset provisions are implemented may be used as a substitute for the awards rate (under the strong assumption that the rate of change is the same for both groups). In either case, the measures across the set of matched counties are combined to form a national estimate of induced entry.

In the matched county design, it is important that the other factors that affect the change in the awards rate over time are the same for both the treatment group and the control group. To illustrate this point, suppose that one county has an economic event that increases their unemployment rate during this period while the matched county does not experience such an event. In this case, both the difference in the unemployment rate and the difference in the benefit provisions will affect the awards rate. It may be impossible to identify the degree to which each of these factors affects the awards rate. In this case, attributing the entire difference to the benefit offset provisions would overstate induced entry.

Tracking Participants. The demonstration does not directly track individuals. It relies on an aggregate county numbers to estimate induced entry. Relying on aggregate county level number can lead to study effects described below.

Required Sample Size. Determining the number of paired counties required in order to detect a 2 percent increase in the awards is a difficult task. This is because the unit of analysis is not the individual, but instead the county. The outcome measure is the aggregate award rate for the county. While using the matched county design has intuitive appeal, the increase in awards of interest (i.e., the effect size) is very small and may be overwhelmed by normal year to year variation or differences across counties that cannot be accounted for under the best matching. As an example of the potential difficulty in measuring induced entry across matched counties, we determined the desired effect size for counties of different sizes, utilizing some “back of the envelope” calculations.

Table 1 shows the number of new awards for a particular year for three sized counties:

	Size Pop. 18-64	Avg. # of new awards	2 % change	5% change	10% change
Small	25,000	75	<2	<4	<8
Medium	100,000	300	6	15	30
Large	200,000	600	12	30	60

Assumption- Average awards rate is ~ 3 per 1000 population 18-64

To be nationally representative, each size county needs to be represented in the sample. Because the change in the number of awards to be identified is very small within the county, it will be extremely difficult to adequately match a sample of counties that will permit estimation of induced entry. Given these small changes in awards, we would guess that a rather large number of matched counties would be required to obtain adequate power.

The quality of the match of counties is critical to the success of this approach. We are not aware of any past study that has successfully matched counties based on the awards rate time trend. We do know that the factors that affected the growth in awards during the 1990s are not well understood. We expect that data on the distribution of health characteristics within a county and the distribution of earnings would have an impact both in determining the county's award rate and the induced entry. However, this data is more relevant to individuals and is not measured in any way that could be adequately employed in a county level analysis. It appears likely that matches will be made with error and the size of the error may be great enough to prevent measurement of induced entry based upon county-level award rates.

Informing Participants. Participants would be informed using a method similar to the one used in a nationally run program (described above). It may be impossible to confine information to the county level resulting in the potential for study effects described below.

Study Effects. The matched counties design will have to deal with the following study effects:

- *Error in the construction of a match*. The results can be very sensitive to the method used to construct the match. Currently, factors that affect a person's decision to apply and be awarded benefits are not well understood. It is possible that the factors that are chosen to perform the match are not the entire set of factors correlated with the person's decision to apply for benefits leading to potential errors in the measurement of induced entry. Second, once the set of factors is determined the match will likely contain some degree of measurement error since it is unlikely that perfect matches can be formed. In both of these cases, the differences in both unobserved and observed factors are likely to change over time at varying rates. As a result, the comparison is unlikely to be a reliable measure of behavior in the absence of the program and the resulting estimate of induced entry can lead to inaccurate conclusions.

- *Exogenous changes over time in the factors affecting awards across the Comparison and Treatment groups*. Even if the entire set of factors that affect awards are identified and if they are perfectly matched prior to the experiment, it is possible for events to occur that affect the factors in one site and not the other. The unemployment rate mentioned in the design section is an example of such a change. See Black, Daniel and Sanders (1998) for a real world example as it relates to the SSDI and SSI program.

  In and out migration. Because the design does not track individuals and relies on aggregate site-level data, the problem of migration into and out of the site can impart an upward bias on the measure of induced entry.

- *Inability to confine information to the county level.* The method used to provide information is likely to have important consequences on induced entry. Information will likely be distributed through mailers, the mass media and outreach efforts. Confining information provided in this manner to a specific geographic location may be impossible. If the information spreads to the comparison site it is possible that the persons in the comparison group will react to benefit offset information rather than existing program information. This will lead to the mis-measurement of induced entry due to the benefit offset provisions.

- *Use of aggregate measures to measure individual behavior.* By aggregating to the "county level", rather than focusing on individuals, with-in site variation in individual level factors such as income levels, health status, etc., that might affect the decision to apply and to be awarded benefits is not identified and considered in the induced entry measure. This results in what is termed aggregation bias and will likely result in the mis-measurement of induced entry due to the benefit offset provisions.

Statistical Issues

- *Evidence on the Magnitude of the Error.* A study by Friedlander and Robins (1995) tested the magnitude of bias that may result from a comparison community design. They were able to take a well-designed random assignment experiment and construct a comparison community design. The results from the randomized experiment design were referred to as the "true effect". They compared the results of the "true effect" to results derived from various constructed comparison groups and found substantial differences in the estimated effect. In a substantial number of cases the estimates from the constructed comparison led to the wrong inference.

  A literature review of matched comparison site studies by Hollister and Hill (1995) also provides examples of a number of comparison site designs that were contaminated by study effects similar to the ones highlighted above.

- *Simulation Results.* A county-level simulation has yet to be performed. Based upon the state level simulations performed, and the assumption that county-level variation in application rates is greater than state-level variation, we believe that such a method would not be successful unless large numbers (100's) of counties were used in the project.

Conclusions. The design will require a large sample and can potentially lead to large program costs, possibly in the hundreds of million dollars. The potential for results to be dramatically different from the "true effect" appears to be large. Results from past matched site studies, based on processes that are likely better understood than the SSDI applications and awards process, have been criticized

based on a poor comparison group. The matched site strategy is likely to produce mis-leading estimates of the impact of induced entry.

Method 3. State Implementation

Design. Under this design, the benefit offset would be implemented in an entire state. A comparison state, or several comparison states, would be selected based on characteristics correlated with the awards decision. The evaluation would consist of comparing the awards rate in the state chosen to that in a comparable state. A model based approach the measure so that inferences on the size of induced entry in a nationally implemented program can be constructed.

Tracking Participants. Individuals are not tracked on an individual basis. The design relies on aggregate state level data.

Sample Size. A large state would be required to measure induced entry. Comparison states would also have to be relatively large. The sample size under this design would be on the order of that shown in Appendix A.

Informing Participants. Participants would be informed in a similar manner as in a nationally implemented program.

Study Effects. The following potential study effects must be considered:

- Measure of induced entry will not be nationally representative.
- Error in the construction of a match. See Method 2.
- Differences in factors affecting the awards rate in each match over time. See Method 2.
- Inability to confine information. Compared to Method 2, this problem may be reduced by using a state.
- In and Out Migration. Compared to Method 2, this problem may be reduced by using a state.

- Aggregation Bias. See Method 2.

Simulation Results In order to assess the ability of a matched state design (using state-level allowance rates) to accurately measure induced entry, a simulation was performed using actual data from matched states with a simulated treatment effect. The results of this simulation appear in Appendix D to this document. This design was shown to perform poorly in measuring a known fixed induced entry effect. The simulation showed that the two-state comparison design could lead to estimates that are of unknown quality and precision and that conclusions based upon such a design could be grossly inaccurate. Statistically speaking, it is unlikely that induced entry could be measured with such a design.

Conclusions. The state design introduces many of the same issues as a matched county design. The use of a state may reduce the in and out migration problem and the inability to confine information, but these problems will still exist.

The design will require a large sample and can potentially lead to large program costs, likely in the hundreds of million dollars. The potential for results to be dramatically different from the "true effect" appears to be large. This is due not only to "study effects", but also due to the inability of the statistical methods to separate the treatment effect from existing cross-state variation. Results from past matched site studies, based on processes that are likely better understood than the SSDI applications and awards process, have been criticized based on a poor comparison group. The matched state demonstration project would be extremely costly and would yield results of unknown accuracy, making them of little value to policymakers.

Method 4. Multiple State Design

Design. This design was developed by the CBO in developing cost estimates for the TWWIIA legislation. Under this design, the benefit offset would be implemented in 5 small states. The CBO actually proposed 10 small states to be included in the induced entry demonstration, 5 for the test of each of two benefit offset proposals. Comparison states

could be selected, or more likely, the demonstration states could be compared to all the remaining states.

The evaluation would consist of comparing the awards rate in the state chosen to that in a comparable state. A model based approach the measure so that inferences on the size of induced entry in a nationally implemented program can be constructed.

Tracking Participants. Individuals are not tracked on an individual basis. The design relies on aggregate state level data.

Sample Size. CBO focussed on using multiple small to measure induced entry, rather than using sample size calculations. The 10 states selected for the two benefit offset demonstration projects represent roughly 3 percent of the U.S. population.

Informing Participants. Participants would be informed in a similar manner as in a nationally implemented program.

Study Effects. The following potential study effects must be considered:

- Measure of induced entry obtained using small states compared to all other (including large states)may not be nationally representative.

- The need to match states is removed, however small states may not be representative.

- Differences in factors affecting the awards rate in demonstration and comparison states over time. See Method 2.

- Inability to confine information. Compared to Method 2, this problem may be reduced by using a state although small states may not be informationally isolated.

- In and Out Migration. Compared to Method 2, this problem may be reduced by using states, though the small states selected by CBO are often small in geographic area and accessible to migrants.

- Aggregation Bias. See Method 2.

Simulation Results In order to assess the ability of the 5 state design (using state-level allowance rates) to accurately measure induced entry, a simulation was performed using actual data from the states identified by the CBO. The results of this simulation appear in Appendix E to this document. This design was shown to perform poorly in measuring a known fixed induced entry effect. The simulation showed that this design could lead to estimates that are of unknown quality and precision and that conclusions based upon such a design could be grossly inaccurate. Statistically speaking, it is unlikely that induced entry could be measured with such a design. An examination of using a 10-state design yielded similar results.

Conclusions. The 5-state design introduces many of the same issues as a matched state and matched county designs. The use of a state may reduce the in and out migration problem and the inability to confine information, but particularly in the geographically small states these problems will still exist.

The 5-state design will require a large sample and can potentially lead to large program costs. In fact, the CBO estimated the cost of the induced entry demonstration to be $190 million dollars, $150 million of which were program costs.[6] The potential for results to be dramatically different from the "true effect" appears to be large. This is due not only to "study effects", but also due to the inability of the statistical methods to separate the treatment effect from existing cross-state variation. The 5-state design demonstration project would be extremely costly and would yield results of unknown accuracy, making them of little value to policymakers.

Method 5. Targeted Population

Design. In this design, the induced entry demonstration is targeted at the population of persons with a severe disability. While the program may induce anyone to apply, in fact, only those with severe disabilities may become

[6] The $190 million cost estimate was for the period 2001 to 2008. The long term costs of testing a benefit offset will be much higher.

allowances. By targeting individuals with a severe disability, a much smaller sample is required to detect a statistically significant change in the awards rate because the prevalence of an award is much larger within this population.

The demonstration design would follow the design outlined under method 1.

Tracking Participants. If sufficient information were provided to participants, demonstration participants have to be tracked to examine application for the program and award status. While the sample to be tracked is large, it is feasible to follow application and awards behavior for participants in the demonstration.

Sample. The sample of persons with severe disabilities first must be identified. It may be possible, to select a sample using those Census respondents required to fill out the long form. If we are unable to use the Census, the design will not be possible because the resources that are required to identify a treatment group and control group will be too large.

Appendix B shows the sample sizes required detect significant induced entry effects. A strong project design will require over **1 million disabled persons**. Even under a design with the smallest acceptable level of statistical power and largest detectable effect, **a sample of 80,000 severely disabled persons for each group is required** (if severely disabled persons could be accurately identified). All other acceptable assumptions require a larger sample.

Information. The issues surrounding the provision of information are the same as those in Method 1. While the required sample sizes are smaller compared to Method 1, they are still large enough to require a large number of resources.

Study Effects. Among other study effects, This design suffers from the major study effect:

- *The target population will not be representative of the population with disabilities.* First, the only survey large enough to identify the necessary sample of persons with severe disabilities is the U. S. CENSUS long form. By the time the benefit offset is implemented, the CENSUS

will be at least two years old. The population identified by the CENSUS will likely change dramatically over this period (deaths, worsening health conditions, etc.). The time lag will also mean that persons who experienced a severe disability after the CENSUS will not be represented in the sample. As a result, the sample will no longer be representative of the population with severe disabilities. This will likely lead to gross mis-measurement of induced entry into the program.

Conclusion. First, the only survey large enough to identify the necessary sample with disabilities is the U.S. CENSUS long form. It is unclear that we will be able to use the CENSUS to identify a sample with severe disabilities for a demonstration. Second, even if we could use the CENSUS, by the time that the benefit offset provisions are implemented the CENSUS will be at least two years old. As a result, the population with severe disabilities that is identified by the CENSUS will not be representative of the existing population with severe disabilities that may apply and be awarded benefits. In fact some may have already applied for benefits, recovered, etc. During the same time period, others may have become disabled who would not be included in the sample.

Given these factors, we do not view the design as reasonable for the purposes of measuring induced entry.

VIII. Alternative Methods to Estimate Induced Entry Using NSHA Data and Other Information

The serious design issues and measurement problems inherent in undertaking a demonstration project to measure the induced entry effect of altering program provisions to include a benefit offset lead us to consider alternatives other than attempts at direct measurement. In addition to the problems listed above, many of the designs discussed provide a relatively poor mechanism for measuring the impact of primary interest: the increase in work and savings generated by the beneficiaries who return to work. Thus not only is the project large and expensive, providing questionable results in the measurement of induced entry, but it also results in a design that yields less definitive results for current beneficiaries.

In lieu of the potential cost and inadequate measurement associated with attempts to "scientifically" measure this complex phenomenon, it is suggested that an alternative method, or in fact methods, be undertaken. This approach will attempt to obtain, and use, the maximum information available to determine the best estimate of the potential for induced entry. The alternative utilizes OCACT estimates, data on potential eligibility from the National Survey of Health and Activity (NSHA-formerly known as the Disability Evaluation Survey), the work response of (and trust fund savings from) current beneficiaries participating in the $1 for $2 project, and a survey to develop estimates of possible induced entry. The multiple approaches will permit us to "triangulate" towards an estimate of induced entry, by examining the sensitivity of estimates to various assumptions among the multiple approaches. Once the work response and trust fund savings have been measured, a point of reference is established and the potential costs of induced entry can be evaluated within the context of this targeted cost level. That is, if small savings are generated among current beneficiaries, only a small level of induced entry can be tolerated and still result in savings or small costs. Thus, this "target" may greatly simplify the analysis of induced entry.

OCACT Estimates

OCACT routinely makes estimates of the impact of policy changes on program size and costs. In fact, OCACT has already produced estimates of the cost of the $1-for-$2 benefit offset. As part of the process of making estimates of the potential inflow, OCACT will prepare estimates using accepted actuarial methods. OCACT will have access to the demonstration data and the NSHA for use in preparing their estimate.

ORES Estimates

In order to make estimates of induced entry and the program costs associated with this entry the ORES estimates must answer the following questions:

1. Who is eligible?
2. Who will be better off under the new program and induced to apply?
3. When will they apply?

4. How much will they work while on the program (i.e., what is the amount of the reduced benefit?

Each of these questions and the information that may be brought to bear are discussed below.

Who is eligible?

The NSHA is a nationally representative study of the prevalence of disability in the general population, with a specific goal of identifying and enumerating those who would qualify for SSDI or SSI benefits if they chose to apply. ORES will use data from the NSHA to determine who is potentially eligible for benefits, and to carefully evaluate the induced entry to the disability rolls that could be anticipated from the $1-for-$2 offset. The NSHA is a well designed, scientifically sound investigation of the prevalence of disabling conditions in the general population. The project will identify disabled individuals, conduct medical examinations, and collect sufficient objective medical information to permit SSA to make simulated disability determinations for these individuals. The determinations will provide SSA with estimates of the number of persons who meet SSA's definition of disability, but are not currently collecting disability benefits. The NSHA, which is currently being pilot tested, will be completed by the time the data from the $1-for-$2 offset is available.

Who will be better off under the $1 for $2 program and induced to apply?

The survey portion of the NSHA will have detailed information with respect to labor force participation and earnings. NSHA eligibility information, along with survey information about labor supply and earnings, will be used to model the incentives and potential behavior for those non-beneficiaries who have been determined in the study to be eligible, based upon medical considerations, for the DI program. Using this information it will be possible to determine not only how many persons would qualify, but also, based upon their earnings and anticipated disability benefit, which individuals would have economic incentives to apply for benefits under the offset provision. This method will require assumptions, or model based estimates, of the labor supply response to any incentives, including labor/leisure tradeoffs.

There will be considerable information that can be brought to bear on making estimates of the impact of the $1 for $2 on those who are eligible. We will know recent earnings as reported in the NSHA survey, be able to calculate SSDI benefit levels based upon matched earnings records and survey reported household composition, and we will know, from the demonstration project, the labor supply and earnings of current beneficiaries provided the $1 for $2 offset. Using this information we can expect to identify several groups of individuals:

1. Those who are clearly much better off entering the rolls and receiving a substantial portion of the DI benefit while working at the current level
2. Those for whom the total income would rise modestly, but might find it advantageous to reduce hours of work or lack health insurance and wish to receive Medicare
3. Those for whom the total income would not change, but might find it advantageous to reduce hours of work or lack health insurance and wish to receive Medicare
4. Those for whom there would be no gain from the benefit offset.

The earnings histories will be used to calculate benefits that would be payable if the individual became eligible for disability benefits. Current earnings would be compared to the total income (earnings + reduced benefit) that would be expected under offset. (This may require behavioral assumptions or modelling. The work response found among current beneficiaries would represent a lower bound on earnings, while current earnings would represent an upper bound.) Also, as a check, it is important to evaluate the incentives under the current program by comparing earnings to the expected benefit plus SGA level earnings. If we find many eligible individuals would be better off under the current program, but have not applied, we may have to consider other approaches. Such a result may indicate a lack of information about DI, uncertainty over eligibility, extreme motivation, etc. that would have to be taken into consideration in predicting who will become a participant.

Who is likely to be induced to enter the program?

This question is more difficult to answer due to the significant behavioral component to the decision to apply.

The potential increase in total income, the desire for more leisure, the likelihood of an extended period of unemployment (or earnings below SGA) necessary to meet the waiting period, level of benefits, wage rate, and other factors will influence this outcome. There is considerable information that can be used in the analysis.

1. Individual earnings histories can be employed to examine variation in earnings that might indicate periods of no, or limited, labor force participation (i.e., years of 0 or low/non-SGA earnings) that would facilitate meeting the 5 month waiting period requirement without altering work patterns.
2. Earnings histories can be examined for trends, such as long-term declines in real earnings which would make benefits more attractive over time.
3. A special survey will be mounted to seek information about disabled person's perceptions of a program with a $1 for $2 relative to the current program providing a heuristic approach to understanding the attractiveness of the offset and seeking the revealed preference of non-beneficiaries.
4. Finally, the work outcomes of the demonstration will give additional insight, as one might expect that the greater the work effort by current beneficiaries the greater the expected magnitude of induced entry (for a number of reasons, including providing an indication of how well the disability decision process excludes individuals capable of work).

<u>When individuals decide to apply</u>

Individuals will decide to apply when it is most advantageous to them. The application process requires a waiting period of 5 months with earnings less than SGA level. Some individuals will be so much better under the offset they will reduce work effort and apply immediately. Others may wait for a period of illness or unemployment to apply. The following information will be utilized to help estimate when the individual will be induced to apply:

- The size of the increase in total income under the benefit offset (a large increase will result in earlier entry)
- The individual's earnings history will be used to detect patterns of years where there are no earnings or earnings

dip near or below SGA facilitating the 5-month waiting period?

- The individual's earnings history will be used to detect patterns trends in real earnings levels, particularly declines that make benefits more attractive over time.

How much will they work while on the program (i.e., what is the amount of the reduced benefit?

Estimating earnings levels of induced entrants are required to determine how much they will work under the benefit offset and the amount of the reduced benefit they will collect. Information on labor supply of the disabled is limited and modelling labor supply parameters based upon beneficiaries or upon eligible non-beneficiaries would likely not provide accurate estimates of labor supply and earnings under an offset. We will, however, know the upper and lower bounds of labor supply of those who will be induced to enter. The labor supply and current earnings reported by eligible non-participants in NSHA will be the upper bound on earnings after entering the rolls. (The matched SS earnings history can be used to determine trends and earnings variation over time for this group.) Using the labor supply current beneficiaries under the $1 for $2 demonstration to model labor supply for induced entrants will provide a lower bound on earnings.

A. Assessing the Stock vs. Flow

This estimation method will not directly address the issue of the short-term effect of induced entry (stock of eligibles) vs. the longer-term induced entry (flow of eligibles), since the NSHA will only directly measure prevalence. The survey data will provide extensive interview and medical evidence with respect to onset. This information may offer the opportunity to make some attempts at breaking out the prevalence rate into the component incidence rates. However, it may be difficult to disentangle the induced entry to get a steady state measure of additional entrants, resulting in an upward bias on the estimate of induced entry.

The process of using existing data sources to estimate the potential for induced entry will probably yield results that are as defensible, or even more defensible, as those which might be obtained from a demonstration project

designed to attempt to measure induced entry. Furthermore, there are none of the ethical considerations of such a project and there is no risk of significant program entry and associated costs of paying benefits to individuals that would not otherwise enter the program. Finally, a project that focuses on beneficiary return to work can be designed to provide better measurement of return to work than would be possible in a project that is designed to also measure induced entry. Ultimately, it is the proper measurement of return to work that is most important since, without a significant increase in return to work, it will not be possible to achieve even cost neutrality, no less program savings.

Survey

The survey approach is an attempt to collect subjective information about individual's perceptions and valuation of the benefit offset relative to the current program. It is strictly a heuristic approach and, on it's own, will not be scientifically valid as individuals often respond differently than they act. The survey could, however, corroborate the results obtained in other analysis.

The goal would be to identify a sample of disabled persons (possibly in NSHA) and determine whether they are working, aware of the disability program, and whether the provision of a benefit offset and basically permanent Medicare would alter their perceptions of the program and decision to apply. There are alternative methods of acquiring information with respect to valuation of the program that may provide more information and allow evaluation of various alternatives, such as different disregard levels, offset rates, etc. Significant background work will need to be performed to develop an adequate survey instrument to undertake this contingent valuation method.

References

Friedlander, Daniel, and Philip K. Robins. "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods." *American Economic Review*, vol. 85, no. 4, September 1995, pp. 923-937.

Hollister, Robinson G. and Jennifer Hill. 1995. "Problems in the Evaluation of Community -Wide Initiatives" in Connell, James P., Anne C. Kubisch, Lisbeth B. Schorr, Carol Weiss eds. *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts* (Washington, D.C.: The Aspen Institute), pp. 127-172.

Appendix A. Power Calculations: Research Sample Sizes to Detect Treatment Effect on Induced Entry

Based upon allowances

Power to detect 2% increase in allowances
A 2% increase in allowances translates to approximately $1billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $400 million)
.80 power requires a sample of 8.9 million persons for each treatment and control group
.90 power requires a sample of 12.3 million persons for each treatment and control group

Power to detect 5% increase in allowances
A 5% increase in allowances translates to approximately $2.6 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $1 billion)
.80 power requires a sample of 1.44 million persons for each treatment and control group
.90 power requires a sample of 2.0 million persons for each treatment and control group

Based upon applications

Power to detect 2% increase in applications
A 2% increase in allowances translates to approximately $1 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $400 million)
.80 power requires a sample of 4.4 million persons for each treatment and control group
.90 power requires a sample of 6.1 million persons for each treatment and control group

Power to detect 5% increase in applications
A 5% increase in allowances translates to approximately $2.6 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $1 billion)
.80 power requires a sample of 720,000 persons for each treatment and control group
.90 power requires a sample of 1.0 million persons for each treatment and control group

Critical Assumptions/Baseline data

U.S. Population 18-64: 154 million
Not at risk: SSDI 4 million
SSI 4.5 million
Population at risk to apply: 145 million
Annual SSDI Applications: 1 million
Annual SSDI Awards: 500,000
Allowance Rates: 50%
Assumes random sampling with no design effect (i.e., no clustering by County, State, etc.)
Alpha=.05

Appendix B. Power Calculations: Research Sample Sizes to Detect Treatment Effect on Induced Entry- Targeting Disabled

Based upon allowances

Power to detect 2% increase in allowances among screened sample – any disability (20%)
A 2% increase in allowances translates to approximately $1 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $400 million)
.80 power requires a sample of 1.4 million persons for each treatment and control group
.90 power requires a sample of 2.0 million persons for each treatment and control group

Power to detect 5% increase in allowances among screened sample – any disability (20%)
A 5% increase in allowances translates to approximately $2.6 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $1 billion)
.80 power requires a sample of 235,000 persons for each treatment and control group
.90 power requires a sample of 325,000 persons for each treatment and control group

Power to detect 2% increase in allowances among screened sample – severe disability (10%)
A 2% increase in allowances translates to approximately $1 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $400 million)
.80 power requires a sample of 500,000 persons for each treatment and control group
.90 power requires a sample of 685,000 million persons for each treatment and control group

Power to detect 5% increase in allowances among screened sample – severe disability (10%)
A 5% increase in allowances translates to approximately $2.6 billion in increased annual DI trust fund expenditures (does not include increased Medicare costs of $1 billion)
.80 power requires a sample of 80,000 persons for each treatment and control group
.90 power requires a sample of 110,000 persons for each treatment and control group

Critical Assumptions/Baseline data

U.S. Population 18-64: 154 million
Not at risk: SSDI 4 million
SSI 4.5 million
Disabled Population at risk to apply:
example 1: Disabled 20% of 154 million= 31 million – 7million beneficiairies=24 million
example 2: Severe Disability 10% of 154 million= 15.4– 7 million beneficiairies=8.4 million
Annual SSDI Applications: 1 million
Annual SSDI Awards: 500,000
20% disabled P1=.02083
P2= .02125 (2% increase); .02188 (5% increase)
10% disabled P1=.05952
P2= .06071 (2% increase); .06250 (5% increase)

Assumes random sampling with no design effect (i.e., no clustering by County, State, etc.)
Alpha=.05

Census: long form with disability questions to 1 in 6 respondents yielding 25.7 million population with long form, possibly 5 million disabled or 2.5 million severely disabled. If response rate =60%, provides 3 million, and 1.5 million respectively, less beneficiaries even these frames of 2.3 million and 800,000 may be too small.

Appendix C. Costs of Conducting a $1 for $2 Induced Entry Demonstration (Costs in Millions of Dollars)

Induced Entry Costs	Maximum DI Program	Probable DI Program	Admin	Medicare
per year cost				
Entry: Min Detectable (2%)	**$62**	**$31**	**$1.2**	**$25**
OCACT Upper (8%)	**$248**	**$124**	**$4.8**	**$100**
5 year demonstration cost				
Entry: Min Detectable (2%)	**$310**	**$155**	**$6**	**$125**
OCACT Upper (8%)	**$1,240**	**$620**	**$24**	**$500**

Assumptions:
Sample size of 9 million persons in treatment test areas
Does not factor in sample design effects that could increase the sample size
Applications cost an average of $1000 through entire process
Maximum program cost is if enterer collect entire average benefit over lifetime
Probable program costs is if the enterer collects one half average benefit over lifetime (However, it is possible that high benefit individuals will be induced to enter and program costs may be considerably higher.)
Lifetime DI Benefit cost for DI worker award: 11.5 years on rolls at average monthly benefit amount
Lifetime Medicare costs for DI Award: $41,000

APPENDIX D

An Evaluation of a Matched State Comparison Design To Measure Induced Entry From The $1-For-$2 Benefit Offset Demonstration Projects

Introduction

Several methods have been proposed as possible ways to measure the induced entry caused by the provision of a $1-for-$2 benefit offset. One suggested method was to identify matched states that could be randomly assigned as a demonstration (treatment) state or used as a comparison state. In the design of the study, comparable states would be matched on key characteristics such as having similar award rates, economic factors, and other characteristics that are believed to influence award rates. The award rates would be monitored over the several years, and along with pre-demonstration awards rates, a statistical model would be employed to test whether differences in the award rates of two states, before and during the demonstration project, indicate that there is an induced entry effect. The purpose of this paper is to simulate an induced entry effect in actual state-level awards data in order to test the ability of the matched state approach, and a differences-in-differences approach, to accurately detect induced entry effects of a size that would be considered meaningful to policymakers.

The Simulation

Data on annual awards for the years 1993 – 1998 were obtained from the Annual Statistical Supplement. Data on population of people between the ages of 18 through 64 for each were obtained from the Statistical Abstract of the United States for the same years in order to compute award rates. Two state pairs, Minnesota -Wisconsin and Indiana-Pennsylvania, were selected for this analysis based on a comparison of their average award rates and informal consideration of other characteristics. This was not intended to test a very refined matching process, but was intended to simply select homogeneous pairs that could be used for an initial test.

The Data

The pairs of states were matched based on award rates (and other known factors) over the years 1993, 1994, and 1995. For the purpose of the simulation, data from 1996 through 1998 was modified to reflect a possible treatment effect in one of the states. We created 3 treatment scenarios: no effect on award rates; a 2% simulated induced entry effect over the period; a 5% induced entry effect; an 8% induced entry effect and a 10% induced entry effect. The first simulation, described above as the "no treatment effect." was based upon the unadulterated awards rates for the pairs of states over the time period. For the 2% simulated induced entry effect, the award rates for one of the states were increased by precisely 2% during years 1996, 1997, and 1998 to simulate a straightforward, fixed 2% treatment effect over those years. Similar data sets were constructed with increases of 5%, 8%, and 10% to the award rates of one of the states for 1996, 1997, and 1998. Then, the simulated increases in awards rates were switched to the other state in the pair to see if there were differences in the estimate of the induced entry depending on which state in the pair received the treatment.

The Model

The model was estimated both using ordinary least squares (OLS) and a logistic regression models. While the OLS results are somewhat easier to interpret, because the award rate is a proportion and bounded between 0 and 1, the logistic regression model is a more appropriate specification. The following equation describes the specification of the LOGIT model:

$$\log\left[\frac{r}{1-r}\right] = b_0 + b_1 x_1 + b_2 x_2 + b_4 x_4 + b_5 x_5 + b_6 x_6 + b_7 x_7 + b_8 x_8 + \varepsilon$$

where r is the award rate,

b_0 is a constant,

b_1 measures the fixed effect for the state with the treatment,

x_1 is a dummy variable representing the state with the treatment for all years,

b_2 measures the treatment effect,

x_2 is a dummy variable for years 1996-98 for the treatment state,

b_4 – b_8 are the coefficients for the specific year effect,
x_4 – x_8 are dummy variables for the years 1994 – 1998,
ε is the error term.

(For the OLS estimate the same model specification was used except the left hand variable was linearly specified as "r", the awards rate)

OLS Results

As described above the model was estimated using ordinary least squares for each pair of states using the following simulated treatment effects: no effect, 2% increase, 5% increase, 8% increase, and 10% increase in awards. Table 1 displays the estimated induced entry effect based upon the actual state data with the induced entry effect simulated in the last 3 years. Those estimates that are based upon coefficients which are statistically significant to the .05 level are denoted with a *. If the estimated does not have a *, then the effect should be interpreted as having no treatment effect detected since it was not statistically different from 0.

Table 1: OLS Estimates of induced entry effect

	Pennsylvania/ Indiana Pair		Minnesota/Wisconsin Pair	
Actual	Treatment: Pennsylvania	Treatment: Indiana	Treatment: Minnesota	Treatment: Wisconsin
No effect	5.2%	-5.2%	-3.4%	3.4%
2%	7.3%	-3.3%	-1.4%	5.4%
5%	10.4%	-0.2%	1.6%	8.4%
8%	13.5%*	2.8%	4.6%	11.4%*
10%	15.5%*	4.8%	6.6%	13.4%*

LOGIT Results

The model was again estimated, this time using the LOGIT technique. The primary purpose was to test how sensitive results are to utilizing different statistical techniques. The LOGIT technique has properties that make it more appropriate for use with bounded data.[7] Table 2 displays the

[7] The OLS estimates are unbiased, but inefficient, meaning the t statistics may not accurately portray the statistical significance of the result. The LOGIT estimate is efficient and consistent.
We felt the OLS estimator would give good results given the fairly tight range of the awards rates, i.e., there was not a wide distribution of awards rates within the bounded range from 0 to 1. We chose to produce both results to demonstrate how consistent and robust the results might be.

estimated induced entry effect based upon the transformed LOGIT coefficient. [8] Those estimates that are based upon coefficients which are statistically significant to the .05 level are denoted with a *. If the estimated does not have a *, then the effect should be interpreted as having no treatment effect detected since it was not statistically different from 0.

Table 2: LOGIT Estimates of induced entry effect

	Pennsylvania/Indiana Pair		Minnesota/Wisconsin Pair	
Actual	Treatment: Pennsylvania	Treatment: Indiana	Treatment: Minnesota	Treatment: Wisconsin
No effect	5.2%	-5.2%	-3.8%	3.8%
2%	7.2%	-3.2%	-1.8%	5.8%
5%	10.1%	-0.3%	1.1%	8.7%
8%	12.9%*	2.5%	3.9%	11.5%*
10%	14.7%*	4.3%	5.8%	13.3%*

The Findings

The results of the simulation show that regardless of the statistical technique (OLS or LOGIT), the results are very consistent and robust with respect to the statistical procedure[9]. For this reason, there is no separate discussion of the two techniques. The findings strongly suggest that measurement error in the estimate of the simulated induced entry effect are the result of underlying data issues, not the choice of statistical method.

The simulations show that, even when states are paired, the choice of which state receives the treatment can have a profound effect on the results that are obtained from the project. In our simulation, the estimated induced entry effect was statistically significant for Pennsylvania when compared to Indiana, but not for Indiana when compared to Pennsylvania. Similarly, Wisconsin was found to have an

[8] The increase in the award rate which is due to the treatment is assessed by computing the slope of the curve with respect to a change in the treatment, given by $\frac{\partial r}{\partial x_2} = b_2 r(1-r)$. A percentage is computed by dividing this number by r and then multiplying by 100. Since the curve is not linear, the slope will vary with the award rate. The average rate over both states and all years was used as the value for r.

[9] Four simulations were found to have statistically significant induced entry effects, and the result was found for both the OLS and LOGIT estimates. The estimates of the size of the induced entry effect are very close. The slight differences in the estimate may occur from the different estimation procedure, or it may have occurred in the conversion of the LOGIT coefficient to a marginal effect and then transforming the marginal effect into a measure of induced entry effect (due to the calculations and rounding error).

induced entry effect relative to Minnesota, but not the reverse. Thus, underlying differences between the states can determine whether a treatment can be detected. This is supported by the estimate in the "no effect" simulation which basically measures state differences in the latter 3 year period, which is not statistically significant but of a magnitude that could be important from a program cost perspective.

The results show estimates of induced entry effect that are statistically significant only in the simulations with Pennsylvania and Wisconsin as the treatment states, and only where the embedded induced entry effects were large (e.g., 8% and 10%). In each case of a significant estimated induced entry effect, the simulated effect was large and the estimated effect overstated the actual effect. In Pennsylvania, for example, the induced entry effect was estimated to be 12.9% and 14.7% (in the LOGIT, even larger in the OLS), both seriously overestimating the embedded induced entry effect of 8% and 10%, respectively. When the treatment is applied to Wisconsin, the effect was statistically significant for only the 8% and 10% simulations, with the measured induced entry being an over-estimate at 11.5% and 13.3%, respectively. For Indiana and Minnesota, the results found consistent (and insignificant) underestimates of the true simulated effect. This result suggests that only large induced entry effects might be found to be significant, and, even when the actual effect is large, it may not produce significant results. And most importantly, the point estimates of the induced entry effect are extremely inaccurate, being over-estimates in some cases underestimates (but insignificant) in others.

It is very troubling that, depending upon which state is assigned to the treatment and which to control, one might obtain an overestimate or an underestimate of the induced entry effect. Thus, even if the awards rate increased by 10%, we could be led to the conclusion that there was no induced entry from the changes in the DI program. An error this large is extremely problematic, as a 10% induced entry effect could mean increased long run program costs of nearly $10 billion per year. Unfortunately, in a real world test, it is impossible to know, either prior to or after the fact, whether the study's state assignment resulted in an induced entry effect that was an over- or under-estimate.

Conclusions

Friedlander and Robins[i] were concerned with the nonexperimental methods used in program evaluations. For their evaluation of the methodology, they chose data on four welfare return-to-work programs – the Arkansas WORK program, the Baltimore Options Program, the San Diego Saturation Work Initiative Model, and the Virginia Employment Services Program. One of the analyses they performed was a two matched state approach. Their conclusion was that "using individuals in one state as a comparison group for individuals in another state can lead to quite inaccurate estimates of the size of a program effect, even if the two groups are matched statistically according to a set of baseline characteristics." Their findings illustrate that "estimates from program effects from cross-state comparisons can be quite far from the true effects." The results of these simulations of induced entry, using actual state level data with imputed fixed treatment effects, are consistent with their findings.

Moffitt[ii] provides a comprehensive review of various approaches that have been proposed to measure induced entry effects. He points out that there has been a lack of attention paid to the effects of changes in a program on induced entry. Most evaluations are based on an examination of the effects of the treatments on those who are presently in the program. Moffitt points out that:

> *"if program entry rates increase as a result of an intervention, and if those who newly enter the rolls have systematically different earnings and labor supply effects than those initially on the rolls, the final average earnings and labor supply effects will be altered. In addition, if a conventional design is used to study the effect of an intervention on program exit rates… the estimates so obtained may also be contaminated by program entry effects. If, again, program entry increases and if those who enter the rolls have different exit rates than those initially on the rolls, final program exit rates will be altered. In both of these cases, proper estimation of earnings, labor supply, and exit-rate effects cannot be obtained in the first place without proper attention to entry-rate effects."*

We find this argument compelling. However, he then points out that conventional experimental methods are not well suited for the analysis of induced entry. Moffitt then

discusses the various nonexperimental methods that are available for addressing the induced entry issue and points to serious shortcomings in every one of them. He concludes the article with a suggestion that a conventional experimental method supplemented by a nonexperimental analysis of induced entry may be the most promising overall strategy.

The simulations performed in this paper, which employ actual state data with embedded fixed treatment effects, lead to the same conclusions as the Moffitt paper. Specifically we found that, as Moffitt suggests, "significant data collection and control for site variables may not be sufficient to absorb all the unobservables on which the cross-site treatment variation is dependent." It seems that our treatment effects were overtaken by unobserved factors, different in each state, which affected the award rates. It is especially problematic if one accepts the OCACT estimate that a $1 for $2 offset (above an SGA level disregard) would result in a 6% increase in awards. CBO indicated that a 6% induced entry effect would be significant and projects should be designed to measure such an increase.[10] It does not appear that the matched state approach is likely to detect such a small difference. Our conclusion, therefore, is that, even though we agree that there is a need for an analysis of the induced entry effect, the two-state comparison method could lead to conclusions that are grossly innacurate and result in estimates of unknown quality and precision. The project required to "measure" induced entry would be extremely costly, and would yield estimates of unknown accuracy, making them of little value to policymakers. Consequently, we believe that alternative methods of assessing induced entry should be pursued.

REFERENCES

[1] Friedlander, D. and Robins, P.K., "Evaluating Program Evaluations: New Evidence on Commonly Used Nonexperimental Methods", *The American Economic Review*, September 1995, pp 923 - 937

[1] Moffitt, R. , "Evaluation Methods for Program Entry Effects", In *Evaluating Welfare and Training Programs*, edited by C. Manski and I. Garfinkel, Cambridge, MA: Harvard University Press

[10] The costs of a 6% increase would be quite high. We estimate that over the long-run, program costs would rise by over $5 billion per year.

APPENDIX A: LOGIT MODEL RESULTS

B. Table 1: Model Estimates for Pennsylvania and Indiana

No Treatment Effect (Pennsylvania)				
	Approx	Approximate 95% Confidence		
Parameter	Estimate	Std Error		Limits
b0	-5.6211	0.0319	-5.7096	-5.5326
b1	-0.0320	0.0321	-0.1211	0.0572
b2	**0.0522**	**0.0446**	**-0.0717**	**0.1761**
b4	-0.0241	0.0398	-0.1347	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	-0.0146	0.0450	-0.1394	0.1103
b7	-0.0109	0.0449	-0.1356	0.1138
b8	0.0541	0.0439	-0.0677	0.1760
Estimated Induced Entry Effect: 5.2%				
2% Increase in Pennsylvania				
	Approx	Approximate 95% Confidence		
Parameter	Estimate	Std Error		Limits
b0	-5.6211	0.0319	-5.7096	-5.5326
b1	-0.0320	0.0321	-0.1211	0.0572
b2	**0.0721**	**0.0444**	**-0.0512**	**0.1954**
b4	-0.0241	0.0398	-0.1347	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	-0.0146	0.0449	-0.1391	0.1100
b7	-0.0109	0.0448	-0.1353	0.1135
b8	0.0542	0.0438	-0.0674	0.1758
Estimated Induced Entry Effect: 7.2%				
5% Increase in Pennsylvania				
	Approx	Approximate 95% Confidence		
Parameter	Estimate	Std Error		Limits
b0	-5.6211	0.0319	-5.7096	-5.5326
b1	-0.0320	0.0321	-0.1211	0.0572
b2	**0.1012**	**0.0441**	**-0.0213**	**0.2237**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	-0.0145	0.0447	-0.1386	0.1096
b7	-0.0110	0.0446	-0.1350	0.1129
b8	0.0542	0.0437	-0.0670	0.1754
Estimated Induced Entry Effect: 10.1%				

8% Increase in Pennsylvania				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.6211	0.0319	-5.7096	-5.5326
b1	-0.0320	0.0321	-0.1211	0.0572
b2	**0.1295**	**0.0439**	**0.00768**	**0.2513**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	-0.0145	0.0445	-0.1382	0.1091
b7	-0.0111	0.0445	-0.1346	0.1124
b8	0.0543	0.0435	-0.0665	0.1751
Estimated Induced Entry Effect: 12.9%				
10% Increase in Pennsylvania				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.6211	0.0319	-5.7096	-5.5326
b1	-0.0320	0.0321	-0.1211	0.0572
b2	**0.1479**	**0.0437**	**0.0266**	**0.2693**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	-0.0145	0.0444	-0.1379	0.1088
b7	-0.0112	0.0444	-0.1344	0.1121
b8	0.0543	0.0434	-0.0662	0.1749
Estimated Induced Entry Effect: 14.7%				

2% Increase in Indiana				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.6531	0.0324	-5.7430	-5.5632
b1	0.0320	0.0321	-0.0572	0.1211
b2	**-0.0324**	**0.0444**	**-0.1557**	**0.0909**
b4	-0.0241	0.0398	-0.1347	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	0.0377	0.0450	-0.0873	0.1626
b7	0.0414	0.0450	-0.0834	0.1662
b8	0.1063	0.0440	-0.0157	0.2284
Estimated Induced Entry Effect: -3.2%				
5% Increase in Indiana				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.6531	0.0324	-5.7430	-5.5632
b1	0.0320	0.0321	-0.0572	0.1211
b2	**-0.00326**	**0.0441**	**-0.1257**	**0.1192**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	0.0376	0.0449	-0.0869	0.1622
b7	0.0415	0.0448	-0.0829	0.1659
b8	0.1063	0.0438	-0.0154	0.2279
Estimated Induced Entry Effect: -0.3%				
8% Increase in Indiana				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.6531	0.0324	-5.7430	-5.5632
b1	0.0320	0.0321	-0.0572	0.1211
b2	**0.0250**	**0.0438**	**-0.0967**	**0.1467**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	0.0376	0.0447	-0.0865	0.1617
b7	0.0416	0.0446	-0.0824	0.1655
b8	0.1062	0.0437	-0.0151	0.2275
Estimated Induced Entry Effect: 2.5%				

10% Increase in Indiana				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.6531	0.0324	-5.7430	-5.5632
b1	0.0320	0.0321	-0.0572	0.1211
b2	**0.0434**	**0.0437**	**-0.0778**	**0.1647**
b4	-0.0241	0.0398	-0.1348	0.0865
b5	0.0275	0.0388	-0.0803	0.1353
b6	0.0376	0.0446	-0.0862	0.1614
b7	0.0416	0.0445	-0.0820	0.1653
b8	0.1062	0.0436	-0.0149	0.2272
Estimated Induced Entry Effect: 4.3%				

Table 2: Model Estimates for Wisconsin and Minnesota

No Treatment Effect (Wisconsin)				
			Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.7627	0.0260	-5.8350	-5.6905
b1	0.0288	0.0261	-0.0438	0.1013
b2	**0.0380**	**0.0381**	**-0.0679**	**0.1438**
b4	-0.00678	0.0316	-0.0945	0.0810
b5	-0.0409	0.0322	-0.1302	0.0484
b6	-0.0754	0.0381	-0.1812	0.0303
b7	-0.1160	0.0387	-0.2234	-0.00861
b8	-0.0936	0.0383	-0.2000	0.0129
Estimated Induced Entry Effect: 3.8%				
2% Increase in Wisconsin				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.7627	0.0261	-5.8353	-5.6901
b1	0.0288	0.0263	-0.0441	0.1017
b2	**0.0579**	**0.0381**	**-0.0480**	**0.1637**
b4	-0.00678	0.0318	-0.0950	0.0814
b5	-0.0409	0.0323	-0.1307	0.0488
b6	-0.0749	0.0382	-0.1808	0.0311
b7	-0.1163	0.0388	-0.2240	-0.00868
b8	-0.0939	0.0384	-0.2006	0.0129
Estimated Induced Entry Effect: 5.8%				
5% Increase in Wisconsin				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.7627	0.0263	-5.8358	-5.6896
b1	0.0288	0.0265	-0.0446	0.1022
b2	**0.0870**	**0.0382**	**-0.0190**	**0.1929**
b4	-0.00678	0.0320	-0.0956	0.0820
b5	-0.0409	0.0326	-0.1313	0.0495
b6	-0.0741	0.0383	-0.1804	0.0322
b7	-0.1168	0.0389	-0.2247	-0.00878
b8	-0.0943	0.0386	-0.2014	0.0128
Estimated Induced Entry Effect: 8.7%				
8% Increase in Wisconsin				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error		Limits
b0	-5.7627	0.0265	-5.8363	-5.6891
b1	0.0288	0.0266	-0.0452	0.1027
b2	**0.1152**	**0.0382**	**0.00927**	**0.2212**
b4	-0.00678	0.0322	-0.0962	0.0827
b5	-0.0409	0.0328	-0.1319	0.0501
b6	-0.0733	0.0384	-0.1799	0.0332
b7	-0.1172	0.0390	-0.2255	-0.00890
b8	-0.0947	0.0387	-0.2021	0.0127
Estimated Induced Entry Effect: 11.5%				

10% Increase in Wisconsin				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.7627	0.0266	-5.8367	-5.6888
b1	0.0288	0.0267	-0.0455	0.1031
b2	**0.1337**	**0.0382**	**0.0277**	**0.2397**
b4	-0.00678	0.0324	-0.0966	0.0830
b5	-0.0409	0.0329	-0.1323	0.0505
b6	-0.0728	0.0385	-0.1796	0.0339
b7	-0.1175	0.0391	-0.2259	-0.00898
b8	-0.0950	0.0388	-0.2026	0.0126
Estimated Induced Entry Effect: 13.3%				
2% Increase in Minnesota				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.7339	0.0258	-5.8055	-5.6624
b1	-0.0288	0.0263	-0.1018	0.0442
b2	**-0.0181**	**0.0381**	**-0.1240**	**0.0878**
b4	-0.00678	0.0318	-0.0950	0.0815
b5	-0.0409	0.0323	-0.1307	0.0489
b6	-0.0380	0.0373	-0.1415	0.0655
b7	-0.0778	0.0379	-0.1829	0.0274
b8	-0.0553	0.0375	-0.1595	0.0489
Estimated Induced Entry Effect: -1.8%				
5% Increase in Minnesota				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.7339	0.0260	-5.8061	-5.6618
b1	-0.0288	0.0265	-0.1024	0.0448
b2	**0.0110**	**0.0382**	**-0.0949**	**0.1169**
b4	-0.00678	0.0320	-0.0958	0.0822
b5	-0.0409	0.0326	-0.1315	0.0496
b6	-0.0388	0.0375	-0.1428	0.0651
b7	-0.0773	0.0380	-0.1829	0.0282
b8	-0.0549	0.0377	-0.1595	0.0498
Estimated Induced Entry Effect: 1.1%				

8% Increase in Minnesota				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.7339	0.0262	-5.8067	-5.6612
b1	-0.0288	0.0267	-0.1029	0.0454
b2	**0.0393**	**0.0382**	**-0.0667**	**0.1453**
b4	-0.00678	0.0323	-0.0965	0.0829
b5	-0.0409	0.0329	-0.1322	0.0503
b6	-0.0396	0.0376	-0.1441	0.0648
b7	-0.0769	0.0382	-0.1829	0.0291
b8	-0.0545	0.0378	-0.1595	0.0506
Estimated Induced Entry Effect: 3.9%				
10% Increase in Minnesota				
		Approx	Approximate 95% Confidence	
Parameter	Estimate	Std Error	Limits	
b0	-5.7339	0.0263	-5.8071	-5.6608
b1	-0.0288	0.0268	-0.1033	0.0457
b2	**0.0577**	**0.0382**	**-0.0483**	**0.1637**
b4	-0.00678	0.0325	-0.0969	0.0834
b5	-0.0409	0.0330	-0.1326	0.0508
b6	-0.0402	0.0377	-0.1449	0.0646
b7	-0.0767	0.0383	-0.1829	0.0296
b8	-0.0542	0.0379	-0.1595	0.0511
Estimated Induced Entry Effect: 5.8%				

APPENDIX E
Induced Entry— Simulation of 5-State Design
(Source:Congressional Budget Office)

Research Design

The CBO constructed a cost estimate for the following two 1-for-2 benefit offset demonstration projects:

1. A benefit offset of $1 for every $2 of earnings above SGA ($500 at the time CBO proposed the research design) with an indefinite benefit offset period—that is, the beneficiary continues to receive the benefit offset for as long as they are medically eligible for the program.

2. A benefit offset of $1 for every $2 of earnings above $85 ($65 SSI earnings disregard + $20 disregard) with an indefinite benefit offset period—that is, the beneficiary continues to receive the benefit offset for as long as they are medically eligible for the program.

The design of the demonstration project was proposed to test the impact, if any, that the two benefit offset projects have on induced entry into and reduced exit from the DI program. CBO assumed that the purpose of the project is to statistically detect an induced entry effect of at least a 6% increase Title II awards and a 79% increase in work by beneficiaries that might result from such a project.

The demonstration design proposed by CBO entailed offering the same benefit offset project to everyone who applies for Title II benefits in five small states so that with two projects a total of 10 states would be included in the demonstration. CBO selected states based upon the total number of awards in 1996—that is, states that made between 1500-3500 awards in 1996 were selected for the analysis. These states include: Delaware, Idaho, Montana, Nebraska, New Hampshire, New Mexico, Rhode Island, South Dakota, Utah, and Vermont. In total, the 10 states made up about 2.9% of national awards in 1996. The estimated total cost of their proposed demonstration is 190 million dollars (155 million dollars in additional benefits paid, and 35 million dollars in contract costs) during the 2001 through 2008 time period.

The purpose of this paper is to investigate whether the CBO research design will result in relatively accurate estimates of the impact of induced entry into the benefit offset program. The paper presents the cost implications based on the estimated results. It then examines whether adjustments to the model can improve the estimates. In the conclusion of the paper, we summarize the general findings and provide a recommendation based upon these findings.

Model

Equation (1) is used to estimate the size of induced entry resulting from a benefit offset provision.

$$y_{s,t} = \alpha_s + \nu_t + \tau\beta + \varepsilon_{s,t} \qquad \textbf{(1)}$$

Where $y_{s,t}$ is the awards rate for each state s at time period t,

α_s represents a state specific effect (state fixed effect),

ν_t represents a year specific effect (year fixed effect),

τ is an indicator that represents the treatment,

β is the coefficient on the treatment (used to construct the estimated treatment effect),

$\varepsilon_{s,t}$ represents the error term,

s indicates each of the 50 states and the District of Columbia,

t indicates each of the 6 years, 1993-1998

Existing annual data for the 50 states and the District of Columbia from 1993 through 1998 was used to estimate the model. The treatment indicator is set equal to 1 for the "treatment states" in years 1996 through 1998. Therefore, we are assuming that *if a treatment was administered to a set of states*, it would be administered on January 1,1996 and continue through 1998. The coefficient (β) is the parameter used to measure the impact of induced entry. It is interpreted as the increase in awards resulting from the treatment, in this case the treatment is the 1-for-2 benefit offset. The coefficient (β) is converted to a percentage by dividing by the average awards rate, which is approximately 0.0039 (3.9 awards per 1000 18-64 year old persons).

In theory, the coefficient (β) on the treatment effect should be equal to 0 in the absence of a simulated effect since no treatment was administered to the "treatment states" during the 1996 to 1998 time period. In practice the coefficient may not be equal to 0 since other factors, such as economic conditions and the administration of the program, that change within the treatment states over time may be included in the treatment coefficient.[11]

An induced entry program effect is simulated in this paper by increasing the awards rates by a constant 6% in each treatment states for each year between 1996 through 1998. This creates an artificial, but consistent, 6% induced entry effect (CBO's target effect size) to be measured by our model. Keep in mind that this is a simulated situation; in practice it is highly unlikely that such a precise increase in awards will occur immediately in each of the treatment states. There might be, for instance, delayed effects of the treatment, different treatment effects in different states, or other factors that affect the time that the treatment effects the treatment states. The results presented in this simulation may be thought of as a "best case scenario" since the impact of the treatment occurs immediately and in the same magnitude in each of the treatment states.

The simulations are intended to provide an indication of the degree of accuracy that might be obtained in the CBO design when a constant and known treatment effect of 6% is simulated. The simulated treatment results are presented in the column entitled "Induced Entry Estimate." The column titled "No Treatment Estimate" is included to show that the coefficient (β) is indeed picking up factors in addition to the treatment that are correlated with the awards rate. It is important to note that, while this "non-treatment" effect can be measured in a simulation, in an actual demonstration project where the treatment is implemented this effect can not be observed since we would not know what would have occurred in the absence of the treatment (i.e., there would be no counterfactual). Therefore, within the context of the demonstration, we would be unable to disentangle the impact of factors other

[11] The model in equation (1) is referred to as a differences-in-differences model. The model attributes changes in the awards rate across the treatment and control states over time to the treatment coefficient (β). In practice, there are factors other than the treatment (e.g., economic conditions and the administration of the program) that may lead to such changes. We discuss the issues associated with attempts to control for these factors below.

than the treatment that are picked up in this coefficient which is purported to measure the treatment effect.

Results of a 5 State Analysis

CBO suggests using 5 states for each experiment. We use the same model shown in equation (1) and divide the set of states into two roughly equivalent groups of five. In Table 1, we used the following treatment states: New Mexico, Rhode Island, South Dakota, Utah, and Vermont.

Table 1. Results of the First Five-State Model

	"Induced Entry Estimate"		"No Treatment Estimate"	
Variable	**Coefficient**	**t-stat**	**Coefficient**	**t-stat**
Treatment	0.00003977	0.46	-.00016788	-1.96
R-squared	0.957		0.958	
Estimated Induced Entry Effect	1.0%		-4.3%	

Note: The estimated model includes time fixed effects and state fixed effects. The model is based on a total of 306 state-time observations. The treatment states include: New Mexico, Rhode Island, South Dakota, Utah, and Vermont.

The first column shows the estimates under a simulated 6% induced entry effect. Under the CBO design for these five states, the coefficient (β) that represents the induced entry effect is not statistically significant and it implies that there is a 1.0% induced entry effect. The estimated impact is not even close to the 6% effect that was actually simulated by the adjustment we made in the data. To further examine the cause of such a gross understatement in the estimate of the treatment effect, we examine the coefficient when the data is not adjusted by the simulated effect, as shown in the column titled "No Treatment Effect". The first column shows that, in the absence of the treatment, the coefficient for the treatment states from 1996 through 1998 has a negative and statistically significant effect on the awards rate. The estimated awards effect ("treatment effect") in the absence of a simulated treatment is calculated to -4.3%. In this case, the treatment variable is clearly picking up factors other than the treatment that are correlated with the awards rate, and that effect has overwhelmed the induced entry effect in the simulation equation.

Table 2 shows the results for the other five states: Delaware, Idaho, Montana, Nebraska, and New Hampshire.

Table 2. Results of the Second Five-State Model

Variable	"Induced Entry Estimate" Coefficient	t-stat.	"No Treatment Estimate" Coefficient	t-stat.
Treatment	0.0002393	2.76	0.0000293	0.34
R-squared	0.956		0.957	
Estimated Induced Entry Effect	6.0%		0.7%	

Note: The estimated model includes time fixed effects and state fixed effects. The model is based on a total of 306 state-time observations. The treatment states include: Delaware, Idaho, Montana, Nebraska, and New Hampshire.

Again, the "Induced Entry Estimate" column shows the estimates under a simulated 6% induced entry effect. In this case, the estimate is statistically significant and it represents a 6% percent increase in benefits. The measured treatment effect appears to be precise, at least in part because, as shown in the second column of Table 2, when the model was estimated without any simulated effect the coefficient was both small and statistically insignificant. The "No Treatment Effect" results indicate that the set of other factors correlated with the treatment tend to be insignificant in this case. While this effect is right on target with the simulated increase in the awards rate, it is important to note that this by no means suggests that these states are the ones to use for an induced entry demonstration project. It is impossible to predict future events that might occur over time in a state which are correlated with the awards rate. While it is possible that eventual demonstration results may be as accurate as those shown in Table 2, it is also possible that the results could look similar to Table 1.

While these two examples show a significant underestimate in one case and a precise estimate in another, any result is possible depending upon whatever changes are occurring among states during the treatment period. It would also be possible, for instance, to obtain results where the estimated induced entry effect may be greater than the true effect of induced entry. Because no true counterfactual would be available during an induced entry demonstration project, it will not be possible to disentangle the true treatment effect from other state/time specific effects, allowing us to accurately identify the impact of other factors on the awards rate.

Results of 10 state Analysis

In order to determine whether a larger demonstration utilizing a greater number of states might be able to better measure induced entry, we also analyzed a design that applies a benefit offset to all 10 of the states identified by the CBO. Table 3 shows the results from the 10-state analysis.

Table 3. Results of 10 State Model

	"Induced Entry Estimate"		"No Treatment Estimate"	
Variable	**Coeff.**	**t-stat**	**Coeff.**	**t-stat.**
Treatment	.0001311	2.02	-.0000777	-1.21
R-squared	0.956		0.957	
Estimated Induced Entry Effect	3.4%		-2.0%	

Note: The estimated model includes time fixed effects and state fixed effects. The model is based on a total of 306 state-time observations. The states include: Delaware, Idaho, Montana, Nebraska, New Hampshire, New Mexico, Rhode Island, South Dakota, Utah, and Vermont.

The "Induced Entry Estimate" column shows that the coefficient on the treatment is 0.000131 and that the coefficient is statistically significant at the 5 percent level. The estimated induced entry effect is 3.4%. In this particular case the 3.4% estimate, while statistically significant, is approximately half the size of the simulated 6% increase in awards rate. We re-iterate that this particular result may or may not occur in the future; actual demonstration results could understate, overstate, or accurately portray the impact of induced entry and we would not know which had occurred. In the present context we know from the results portrayed in the "No Treatment Estimate" in column 2 above, that the coefficient was not statistically significant, but was large enough to ultimately reduce the estimate of induced entry. The estimate of induced entry has been affected by some set of state and time factors (other than the treatment) that are correlated with the awards.

Consequences of Results

To illustrate the potential consequences associated with inaccurate estimates of induced entry effects, we address the issue of most interest to policymakers: the estimate of

costs associated with induced entry. To demonstrate the potential error in the cost estimates that would be provided to policymakers, we compare the cost estimates that would result from each of the three estimated induced entry effects and compare it to the "true effect" (6% increase in awards) that we simulate in this case. In Table 4 we present the CBO estimate of the impact on costs of a 6% induced entry effect. We then show the cost estimates that result from each of the estimated models[12].

Table 4. Costs of Induced Entry in Nationally Implemented 1-for-2 Benefit Offset Program, 2001-2008

Estimated Induced Entry Effect	6% Induced Entry Cost (in Billions)	Estimated Induced Entry Cost (in Billions)	Absolute Difference in Cost Estimate (in Billions)
Table 1 Estimate	$5.35	$0.89	$4.46
Table 2 Estimate	$5.35	$5.35	0
Table 3 Estimate	$5.35	$3.03	$2.32

Source: Authors' calculations from CBO cost estimates.

The table shows that the understated induced entry effect shown in Table 1 and Table 3 can have very important implications on the Cost estimate. The 5-state estimate shown in Table 1 leads to a $4.46 billion (83%) underestimate of the true cost of the program! The results from Table 3 show a $2.32 billion (43%) underestimate of the cost of the program. These differences are substantial. While the Table 2 results appear to be right on target, as discussed above, there is no certainty that if the demonstration were conducted today that the same result would occur as it is not possible predict how

12 The estimates in Table 4 were constructed as follows. The demonstration project states make up approximately 2.9% of annual awards and were estimated by CBO to cost 155 million dollars in benefits from 2001 through 2008. To get to the national estimate shown in the column titled "6% Induced Entry Cost" we multiplied the 155 million dollar number by 34.48 (100/2.9) and arrived at a cost on a national level of 5.35 billion dollars. The column titled "Estimated Induced Entry Cost" is based upon a proportionate allocation of costs- constructed by dividing the estimated induced entry effect (shown in Tables 1, 2 and 3) by the 6% induced entry effect and then multiplying by the $5.35 billion cost of a 6% induced entry effect. The last column is simply the absolute value of the difference between these two effects.

factors associated with the awards rate might change over time.

Can these Estimates Be Improved?

It may be possible to assess the influence of other factors associated with the awards rate on the treatment coefficient. The first step in the assessment is to specify factors that may be correlated with the awards rate before the experiment takes place. Then, one can examine how these factors change over time in the treatment states compared to the remaining states. If it appears as though one or more of these factors change at a different rate both over time and across treatment and comparison sites, then it may be possible to assess whether certain states might be excluded from the analysis, and to determine whether one might anticipate that estimate either overstates or understates the true induced entry effect. However, such decisions would be subjective and only qualitative information would be obtained from this exercise since we would not know the extent to which any given factor influences award rates or how the factors interact.

In some cases it may be possible to include other factors into a model to improve the estimate of the treatment effect. However, including other factors into a model requires a great deal of information on the process by which the factors affect the outcome of interest.[13] In many cases, factors are included in a more-or-less "ad hoc" fashion after the experiment has been performed, without careful consideration of the true process by which the factor interacts with the treatment and/or effects the outcome of interest.[14] Including factors in an "ad hoc" fashion tends to result in a mis-specified model. A mis-specified model usually results in an estimated impact of the treatment that is no better, and in some cases could be worse, than the estimated impact of the treatment when the factors aren't included. For the DI program, there are several factors that affect the state awards rate, including: state level economic conditions, characteristics of individuals who live within the state, the state level administration of the DI program, etc. The manner in which

[13] Models that carefully take into account the process by which factors affect the outcome of interest are classified as "structural models".

[14] These are typically classified as "reduced form" models. See Lucas (1976) for a discussion of the issues surrounding using "reduced form" models for econometric policy evaluation.

these factors interact with one another, interact with the treatment and ultimately affect the state awards rate is unknown. As a result, these factors could only be included in the model in an "ad hoc" manner. It is likely that the resulting model will be mis-specified and that the estimated impact of the benefit offset is no better, and in fact could be worse, than the estimated impact in a model that does not include these factors.

Conclusion

This paper performed simulations based on a research design for a state level demonstration project to measure induced entry that was proposed by CBO. It shows that under two of the CBO designs, the estimated impact of induced entry would be way off the mark if the benefit offset had been implemented in 1996 and if the true induced entry effect was 6%. The consequences of the understated estimates are of great concern since policymakers would have been misled in terms of the estimated costs of induced entry that would occur if this were implemented nationally. For example, if the 10-state model were used, the cost estimate would be $2.32 billion (about 43%) below the true induced entry cost.

The results presented in this paper are consistent with the literature on research designs that use small numbers of geographic units to assess a treatment effect. [15] That literature warns that the ability to accurately detect effects with any level of certainty is often compromised by these types of designs. For the evaluation of the DI program, the problems with geographic location based designs are much greater for the following three reasons:

1. **The size of an induced entry effect that has important policy implications is extremely small and difficult to detect in the general population.** As a result, it is

[15] The literature on the issues with geographic location based designs is summarized in Hollister and Hill (1996) who conclude, based on their review of the literature, that:

"First, the few existing studies of the problem show that the magnitude of errors in inference can be quite substantial even when the most sophisticated methods are used. Second, the bias can be in either direction: we may not only be led to conclude that an intervention has had what we consider to be positive impacts when in fact it had none, we may also find ourselves confronted with impact estimates which indicate, due to bias, that the intervention was actually harmful; we may be misled either to promote policies which in fact use up resources and provide few benefits or we may be led to discard types of interventions as unsuccessful which actually have underlying merit. (Hollister and Hill, p. 27)"

very difficult to separate out an induced entry effect from random variation in the awards rate.

2. **There is a greater deal of heterogeneity within the population with disabilities relative to the population studied in other Social programs (e.g. AFDC mothers).** As a result, the factors that affect a person's decision to apply for DI and be awarded benefits are very different from person to person and it is very difficult to summarize disabled worker induced entry behavior in a simple statistical model.

3. **The factors correlated with the DI awards rate are not clearly understood.** Techniques that might improve a geographic site based research design, such as matching, require a reasonable degree of information on the factors that explain the changes in DI awards over time. For the DI program, the information on factors that explain the changes in DI awards over time is very limited at best. Therefore, it is unclear whether techniques, such as matching, would lead to improved estimates.

Therefore, the standard methods that might be used to obtain estimates of program effects are more likely to be problematic in an evaluation of the DI program than in other applications.

Perhaps of greatest concern is the fact that the simulations presented in this paper are based on a simulated induced entry effect that should be relatively easy to identify. That is, it is based on a consistent and precise 6% increase in awards that occurs in all treatment states at the same point in time and for the exact same duration. It is highly unlikely that this "best case scenario" will occur in real world circumstances, such as during the implementation of the demonstration project. It is likely that the true effect on the awards rate will vary across treatment states and may change over time. In practice, this makes the precise identification of the treatment effect much more difficult compared to the effect simulated in this paper. As a result, in the real world our ability to produce accurate estimates of the induced entry effect will be subject to a much greater level of uncertainty.

The conclusion that we draw from the results in this analysis, as well as the existing literature on geographic

site-based research designs, is that the induced entry effect estimated based upon a geographic site-based design is likely to be very different than the true induced entry effect. How close the estimate is to the true effect, and whether the estimate is larger or smaller than the true effect, will not be known. The results in this simulation are strictly empirical results, and do not address any of the issues about a demonstration project's ability to create and measure the behavior that is of interest, in this case induced entry. We strongly recommend that policymakers carefully consider this evidence prior to implementing a geographic site-based research design that will cost $190 million over a 7-year period, and additional program costs into the foreseeable future. The $190 million figure is a large sum of money to pay for an estimate of induced entry of unknown accuracy, and potentially erroneous information that could lead policymakers to act to implement the policy on a national level leading to substantially higher program costs (Table 4).

References

Hollister, Robinson G. and Jennifer Hill. 1995. "Problems in the Evaluation of Community -Wide Initiatives" in Connell, James P., Anne C. Kubisch, Lisbeth B. Schorr, Carol Weiss eds. *New Approaches to Evaluating Community Initiatives: Concepts, Methods and Contexts* (Washington, D.C.: The Aspen Institute), pp. 127-172.

Lucas Jr., Robert E. 1976. "Econometric Policy Evaluation: A Critique," *Journal of Monetary Economics*, 1(2) Supplementary Series 1976, pages 19-46.