CENTER for RETIREMENT RESEARCH at BOSTON COLLEGE

COMPARATIVE REGRESSION DISCONTINUITY AND REGRESSION DISCONTINUITY AS ALTERNATIVES TO RANDOMIZED CONTROLLED TRIALS FOR ESTIMATING AVERAGE TREATMENT EFFECTS: EVIDENCE FROM THE BENEFIT OFFSET NATIONAL DEMONSTRATION

Duncan Chaplin, Charles Tilley, Denise Hoffman, and John T. Jones

CRR WP 2022-7
August 2022

Center for Retirement Research at Boston College
Haley House
140 Commonwealth Avenue
Chestnut Hill, MA 02467
Tel: 617-552-1762 Fax: 617-552-0191
https://crr.bc.edu

Duncan Chaplin is a principal researcher at Mathematica. Charles Tilley is a senior analyst at Mathematica. Denise Hoffman is a senior researcher at Mathematica. John Jones is an economist with the U.S. Social Security Administration's Office of Research, Demonstration, and Employment Support, Office of Retirement and Disability Policy. The research reported herein was pursuant to a grant from the U.S. Social Security Administration (SSA) funded as part of the Retirement and Disability Research Consortium. The opinions and conclusions expressed are solely those of the authors and do not represent the opinions or policy of SSA, any agency of the federal government, Mathematica, or Boston College. Neither the United States Government nor any agency thereof, nor any of their employees, make any warranty, express or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of the contents of this report. Reference herein to any specific commercial product, process or service by trade name, trademark, manufacturer, or otherwise does not necessarily constitute or imply endorsement, recommendation or favoring by the United States Government or any agency thereof. The authors would like to thank John Deke at Mathematica for carefully reviewing this work and guiding us through the process. They would like to thank Dave Stapleton at Mathematica for helping them hone this idea. They also received valuable advice from Phil Gleason, Peter Schochet, Hanley Chiang, Natalya Verbitsky, and many other colleagues at Mathematica as well as from external reviewers including Austin Nicolas (Abt Associates), Lakshmi K. Raut (Social Security Administration), and Richard Chard (Social Security Administration). Serge Lukashanets helped with programming and Chelsea Poshkus helped with project management. The authors also received helpful advice from Thomas Cook (Northwestern) and Heinrich Hock (AIR).

About the Center for Retirement Research

The Center for Retirement Research at Boston College, part of a consortium that includes parallel centers at the National Bureau of Economic Research, the University of Michigan, and the University of Wisconsin-Madison, was established in 1998 through a grant from the U.S. Social Security Administration. The Center's mission is to produce first-class research and forge a strong link between the academic community and decision-makers in the public and private sectors around an issue of critical importance to the nation's future. To achieve this mission, the Center conducts a wide variety of research projects, transmits new findings to a broad audience, trains new scholars, and broadens access to valuable data sources.

Center for Retirement Research at Boston College
Haley House
140 Commonwealth Ave
Chestnut Hill, MA 02467
Tel: 617-552-1762 Fax: 617-552-0191
https://crr.bc.edu/

Affiliated Institutions:
The Brookings Institution
Mathematica – Center for Studying Disability Policy
Syracuse University
Urban Institute

Abstract

In this paper we use data from an evaluation of the Benefit Offset National Demonstration (BOND) to evaluate the efficacy of using comparative regression discontinuity (CRD) and regression discontinuity (RD) relative to a randomized controlled trial (RCT). BOND is a large demonstration intended to promote return to work among people with disabilities who receive Social Security Disability Insurance (DI). RD is known as a relatively rigorous non-experimental method but produces imprecise results that apply to small populations. CRD is a promising enhancement that addresses these issues. The CRD and RD methods are potentially attractive because they can be used in contexts in which RCTs are challenging or infeasible. However, the bias of findings from CRD and RD studies is unknown in the context of DI. In this paper, we estimate CRD and RD models using simulated assignment to the BOND treatment group based on the duration of DI receipt at the start of BOND. We compare the CRD and RD estimates to RCT estimates. While the findings are not intended to revise the well-established evidence evaluating BOND, they can be used to help interpret the results from CRD and RD studies on other income support interventions for people with disabilities and to inform future study designs.

Our paper has two key limitations. First, our RD models are far from ideal. This limits the degree to which our RD results generalize to what would be found with state-of-the art RD models. Second, our results may not generalize to other populations. Our analysis was based on BOND beneficiaries who were representative of the larger DI population at the time of BOND random assignment but may not reflect the DI population in more recent years.

The paper found that:

- Average bias from CRD and RD is generally below 0.02 standard deviations in absolute size for the groups of bias estimates we analyzed.
- Given the precision that may be needed to evaluate interventions like BOND, the standard deviation of bias (after accounting for sampling error) is nontrivial, generally between 0.02 and 0.07 standard deviations for the groups of bias estimates we analyzed.

The policy implications of the findings are:

- When designing and interpreting results from CRD and RD evaluations, it is important to note that both produce biased estimates suggesting that their results be interpreted with more caution than those from an RCT with similar standard errors.
- This bias appears to be larger in the presence of major non-linearities in the relationship between the running variable and the lagged outcome for CRD.

Introduction

There are numerous methods for evaluating the treatment effects of a policy intervention. Three of these are regression discontinuity (RD), an enhanced version of it, comparative regression discontinuity (CRD), and randomized control trials (RCTs). RD can be used when treatment is allocated based on a cut point on a variable measured before the treatment happens. RD estimates impacts at that cut point. CRD adds additional information to an RD model to facilitate estimating impacts away from the cut point. RCTs can only be used when treatment is assigned randomly. In all cases, impacts are estimated by comparing outcomes of the treated group to those who were not treated, though adjustments are needed in the cases of RD and CRD. In some cases, practitioners have sufficient freedom in designing a given policy intervention to make it possible to use any of these evaluation methods. In other cases, some methods may not be possible because of legal or practical barriers to assigning people to the treatment group. Regardless, it is useful to know the efficacy of each method in estimating the average treatment effects. It would help in the design of the study of a policy intervention (when possible) or help in interpreting the results, if the study design cannot be adjusted.

That is the objective of this paper – to compare the efficacy of RD, CRD, and RCT methods in estimating the average treatment effects of a policy intervention. We do this within the context of the Social Security Disability Insurance (DI) program. The results can be used when making decisions about whether to use an RCT model rather than an RD or CRD and when interpreting RD and CRD results. RD is well known for being a relatively rigorous non-experimental method and CRD is a promising potential enhancement (Tang et al. 2017). Both can expand opportunities to evaluate programs and policies by avoiding some of the challenges of RCTs. Although RCTs are known to produce unbiased impact estimates with minimal assumptions, they are often difficult to implement. For example, in a retrospective study where program participants were selected using a cut point on a continuous variable, an RCT would not be feasible because assignment happened prior to the study being designed, but RD or CRD analyses could be used. Even in a prospective study, an RCT might not be feasible if there is a legal requirement that all eligible people are served; but, if one of the eligibility criteria is a cut point on a continuous variable, then RD or CRD could be feasible.

The RCT, RD, and CRD methods differ substantially in the target populations, how individuals are assigned to treatment status, and how impacts are estimated. In an RCT

researchers can obtain an unbiased estimate of the average treatment effect (ATE) by comparing outcomes of individuals who were randomly assigned to the treatment group with outcomes of those who were assigned to the control group. In an RD model treatment status is determined entirely by a cut point on a pre-treatment variable known as a running variable. Individuals on one side of the cut point are treated and those on the other side are not treated. RD researchers can obtain a consistent estimate of the impact at the cut point by comparing outcomes for individuals in the treatment group to those in the comparison group. To reduce the potential for bias due to differences in the running variable, the analysis sample is constrained to be close to the cut point and regression adjustment is often used for any remaining differences. CRD is similar to RD but is designed to facilitate estimating an ATE instead of just an impact at the cut point. It does this by incorporating data on individuals with values farther from the cut point and by adding in an additional outcome variable where no impacts are expected, for example a lagged outcome.[1] An ATE can be estimated by comparing outcomes of the treatment group with those of the comparison group, adjusting for differences in the values of the running variable and the lagged outcome.

The performance of RD versus RCTs has been studied in several contexts and at least 15 times (Chaplin et al. 2018). Most of these studies focused on education or politics, while one analyzed the impacts of a welfare program (Wing and Cook 2013). Few studies have tested the efficacy of CRD; all those studies have all been in education (Kisbu-Sakarya et al. 2018; Tang et al. 2018; Tang et al. 2017) or health (Wing and Cook 2013). However, none of these studies have investigated RD or CRD in the context of income support programs for people with disabilities.

This paper analyzes the efficacy of RD and CRD in a specific policy context. In particular, we focus on how these methods affect estimated impacts of an innovation that was assigned to a small randomly chosen treatment group of potential DI recipients in a selection of sites. DI is the United States' largest income support program for people with disabilities. It provides cash benefits for people unable to engage in substantial gainful activity because of long-lasting physical or mental impairments. In 2018, the Social Security Administration (SSA),

[1] CRD models can also use outcome variables that cover both post- and pre-treatment periods as long as no impacts are expected for those variables (Tang et al 2017).

which oversees DI, paid benefits to almost 10 million DI beneficiaries with disabilities (SSA 2019).[2]

The innovation covered in this paper is SSA's largest demonstration to date: the Benefit Offset National Demonstration (BOND). SSA has conducted seven large-scale RCT demonstrations to test innovations to the DI program over the past nearly two decades. BOND, which included nearly 1 million beneficiaries in the program's evaluation, changed the way earnings affected DI cash benefit payments. BOND control subjects were and still are governed by the laws current at the time of the BOND evaluation. Under those laws SSA withheld the entire monthly benefit for a DI beneficiary if, after a 12-month period to test work, their earnings exceeded a programmatic threshold set annually.[3] BOND treatment enrollees were instead subject to a $1 reduction in benefits for every $2 in earnings above the programmatic threshold. The BOND innovations had the potential to simultaneously increase earnings and decrease total DI benefit payments. The evaluation of BOND found no statistically significant increases in earnings and yielded an increase, rather than a reduction in DI benefits (Gubits et al. 2018). The demonstration and its evaluation were costly to implement from an administrative perspective but saved the DI trust fund money relative to implementing the innovation without first testing it, given the increase in average DI benefit payments that was found. In this analysis, we investigate how biased results from an evaluation of BOND might have been had it used the CRD or RD method instead of an RCT. The results are not of interest for assessing the BOND program since the program has already been evaluated using an RCT, but may be useful when interpreting results regarding other DI program innovations that have not been evaluated using RCTs and when making decisions about how those innovations might be evaluated.

CRD and RD are related and often feasible for retrospective evaluations of DI and related programs because some rules governing eligibility and benefits rely on time-based variables such as age and benefit duration. For example, Chen and van der Klaauw (2008) used an age-based RD to evaluate the impact of DI on labor supply and Deshpande (2016) used an age-based RD to evaluate the effect of removing Supplemental Security Income (SSI) eligibility on earnings and

[2] In 2018, SSA made DI payments to 8,537,332 disabled workers, 254,581 disabled widow(er)s, and 1,127,181 disabled adult children.

[3] Under that law, SSA withheld benefits for months in which beneficiaries engaged in substantial gainful activity. Non-blind beneficiaries who earned more than $1,310 in a month and blind beneficiaries who earned more than $2,190 in a month were considered to have engaged in substantial gainful activity.

income. Similarly, while not using a time-based variable, Gelber et al. (2015, 2017) used a regression kink design, which is very similar to RD, to estimate the effect of DI on mortality and earnings.

We use the duration of DI benefit receipt as the key eligibility criterion (running variable).[4] Thus, we are simulating what might happen if the BOND program was implemented using months of DI receipt as the eligibility criteria and we were to estimate impacts using RD and CRD in that scenario. SSA may choose to use other variables to determine eligibility for future program innovations (for example, age, baseline income, health status, etc.). We chose duration of DI benefit receipt in part because it is a variable that is associated with later outcomes and takes on a large number of unique values (needed to obtain reasonably precise impact estimates). We chose DI duration over the other possible running variables (age, income, and health status) because duration was expected to lead to differential impacts of BOND and the BOND sample was stratified according to duration to ensure an oversample of short-duration beneficiaries. Furthermore, the experiences of short-duration beneficiaries are indicative of how new DI awardees would experience any new policy that became law (Stapleton et al 2010).

In future research it may be worth exploring alternative running variables. The relationship between the running variable we chose (DI duration) and the outcomes is fairly linear, as shown in Appendix C. If it was even more linear for the other possible running variables then we might expect less bias for them.

We focus on the first few years of DI benefit receipt for several reasons. First, return to substantial employment is most likely within the first five years of DI benefit receipt (Liu and Stapleton 2011). Indeed, BOND was designed to oversample beneficiaries who received DI for three years or fewer in part because of the association between short duration of DI benefit receipt and positive employment outcomes (Bell et al. 2011). Second, offering a policy to new awardees helps support long-term projections of potential impacts for future beneficiaries, for whom the policy would presumably be available upon DI entry. Finally, as a practical matter, beneficiaries who received DI benefits for three years or fewer comprised about 50 percent of the

[4] Following the BOND evaluation, we used the disability adjudication date as the start of the duration of DI benefit receipt and, when missing, used the date of DI entitlement, with the former rounded to the nearest month. Entitlement dates are always the first of a month.

BOND evaluation sample, leading to large counts of beneficiaries who had received DI benefits for only a few years. This enables us to detect relatively small amounts of bias.

This work can be used to help guide future research on income support programs for people with disabilities. More precisely, it can be used to help decide whether or not to do a prospective RCT and for interpreting RD and CRD results. As noted above, SSA has already conducted seven large-scale RCT demonstrations to test innovations to the DI program. The results in this paper would not be relevant to any of those innovations since the RCTs were already conducted. However, our results could be used to help inform decisions about whether or not an RCT is needed when testing future innovations. If RD or CRD could be used in place of RCTs, then this could save significant evaluation resources, especially for recruiting sites to participate in a study. Even if the program and data collection costs are not affected, getting policy-makers to agree to an RCT can be difficult (Stuart et al 2017, McCann 2019, and Tipton et al 2021) and thus, possibly more expensive than getting them to agree to continuing or starting to use a cut point value to allocate an innovation. Our results could also help readers interpret RD and CRD results, and specifically to help provide readers with an understanding of the potential for bias in such studies.

Overview of Within-Study Comparison

In this paper, we conduct a within-study comparison. Within-study comparisons are designed to estimate bias by comparing results from one type of statistical model to another more rigorous one, with the latter normally being an RCT (Wong and Steiner 2018). An early example includes work by Lalonde (1986), while more recent efforts include that of Chaplin et al (2018) and Weidmann and Miratrix (2021). Cook, Shadish and Wong (2008) describe key features of within-study comparisons.

In this study, we estimate bias obtained when estimating impacts using CRD or RD compared to what one would obtain when using a closely related RCT. We produce the CRD and RD estimates by first creating a quasi-experimental design from the data generated for the RCT. To do this we first identify a running variable. We then assign treated observations from the RCT that are on one side of a selected cut point value of the running variable as being in the synthetic treatment group and assign observations not treated in the RCT that are on the other side of the cut point to the synthetic control group. In this subset of the data we estimate an RD

regression. This is a common approach to investigating RD designs in within-study comparisons (WSCs) as in, for example, Gleason et al (2018). The approach requires a large RCT because the data requirements are substantial.

Our estimates of bias are fairly standard for the CRD models since the CRD and RCT models we use are designed to estimate the same estimand (an ATE). They can be thought of as capturing only internal validity bias. The story is more complicated for RD. In that case, the RD models estimate impacts at the cut point but the RCT models estimate impacts for populations that include the cut point but that also include at least some observations away from the cut point.[5] Since treatment effects can vary away from the cut point, these estimands can differ. We still refer to this as bias because RD estimates are generally used to draw conclusions about populations that include observations away from the cut point. Were this not the case then RD estimates would not be relevant for a noticeable fraction of any population. In other words, the bias we are estimating for RD incorporates both internal and external validity bias (Olsen et al 2018).

Sources of Variation in Bias Estimates and Groups of Bias Estimates

While our paper is based on the BOND study, our goal is to estimate a distribution of bias that might be found in future studies. To generate this distribution we create 3,600 bias estimates that vary by method (RD vs CRD), estimand, cut point, outcome, demographics (age and gender), side of the cut point the treatment group is on, whether covariates are used, and how many months of data of the running variable are included. We then break these bias estimates up into sets of non-overlapping groups. Most sets have two groups but some have more. We then test to see how estimated average bias and the estimated variation in bias differs from 0 and how the estimated variation in bias differs across the groups of bias estimates within each set.[6] Each set has groups that differ based on one or two of the characteristics used to create the bias estimates (estimand, cut point, outcome, demographics, etc.). However, the groups in each set are perfectly balanced with each other on the other characteristics. Thus, when we compare—for

[5] This is based on the continuity framework for interpreting RD results. More recently some authors have proposed using a local randomization framework which is based on the assumption that observations near the cut point can be treated as if they were randomly assigned to treatment status (Cattaneo, Frandsen, and Titiunik 2022).

[6] Since we estimate 3,600 bias estimates one might expect 5 percent to be statistically significant at the 5 percent level due to chance. We avoid this issue by focusing on estimating the mean and variation of bias.

example—the group of RD bias estimates with the group of CRD bias estimates there is variation across bias estimates within the RD and CRD groups caused by each of the other characteristics described above. However, each RD bias estimate matches to one CRD bias estimate on each of those characteristics. A similar point holds for each of the other sets of non-overlapping groups.

We generate variation in bias estimates by estimand by using four bandwidths in the CRD and RCT models. For each bandwidth we estimate impacts for beneficiaries with DI durations no more than the bandwidth from the running variable cut point and on the same side of the cut point as the synthetic treatment assignment. The bandwidths include (i) 0.5 years (ii) two years, (iii) four years, and (iv) all possible years. The magnitude of bias might be expected to increase as the bandwidth gets larger, especially for the RD model which estimates impacts using a 0.5 year bandwidth regardless of the bandwidth being used in the RCT and CRD models.

We generate variation in bias estimates by cut point by using four cut points chosen to target potential DI policies—two, three, four, and five years on DI as of June 1, 2011 (the approximate mid-point of notification of assignment to BOND).[7] We have no a priori reasons to believe that the magnitude of bias will differ across the bias estimates based on this characteristic or most of the others considered in this study, but we also have no reason to expect that the levels of bias would be similar, so these additional sources of variation can help to estimate the potential for variation in bias.

We estimate bias using five outcomes from 2014, the third full year after BOND random assignment. Our outcomes include earnings, employment, earned BOND yearly amount (the programmatic threshold discussed earlier), months with DI benefits, and annual DI benefits due.[8] We have lagged measures for each outcome, corresponding to the same variable measured in 2010, a few months before BOND random assignment in May 2011. We standardize each outcome and lagged outcome to make the results comparable in magnitude. We also multiply the signs of the months of DI and annual DI benefits variables by negative one so that positive

[7] BOND random assignment occurred in May 2011 and treatment subjects were notified of their assignment between May and August 2011. Notification was initially slated to end in July 2011.

[8] Earned BOND yearly amount is set to 1 if annual earnings are above 12 times a monthly DI programmatic threshold used to determine initial and ongoing eligibility for DI and 0 otherwise. In 2014, the threshold was $1,070 per month for non-blind beneficiaries, which translated to $12,840 annually.

bias suggests that the RD or CRD model appears to be biased in favor of the intervention relative to the RCT estimate, regardless of the outcome.

We estimate bias for five populations of individuals that differ based on their demographic characteristics. These include four populations created by splitting the data into approximately evenly-sized pieces based on gender and age (above and below median age) and a fifth equal to the total population.

We estimate bias on both sides of the cut points. More precisely, we estimate impacts assuming the treatment group is above the cut point on the running variable, and then, assuming the treatment group is below the cut point. Although we presume that the main policy of interest is in delivering services to those with durations below a given cut point, this approach of estimating bias on both sides of the cut point has the methodological advantage of doubling the sample size of the bias estimates.

Specification tests have been developed to help identify CRD models that might produce biased estimates (Tang et al 2017). We group the bias estimates based on the results of CRD specification tests for discontinuities at the cut point; differences in slopes above and below the cut points; and differences in slopes between the outcome and lagged outcome. We also consider various combinations of these tests giving us 5 sets of non-overlapping groups. In this case we would expect more bias when the tests for bias reject the null of no bias.

We estimate bias with and without covariates. We include models without covariates in part because covariates are not needed for RD and may not be needed for CRD. In addition, as discussed later, we use aggregate data and this limits our ability to estimate models with covariates.

Finally, we estimate our models using two samples of data—one that uses all data for beneficiaries with from 4 months to 150 months of DI eligibility at the start of the study, and the other going from 15 to 150 months. We do this because of a large nonlinearity in lagged outcomes observed around 15 months (see Appendix E).

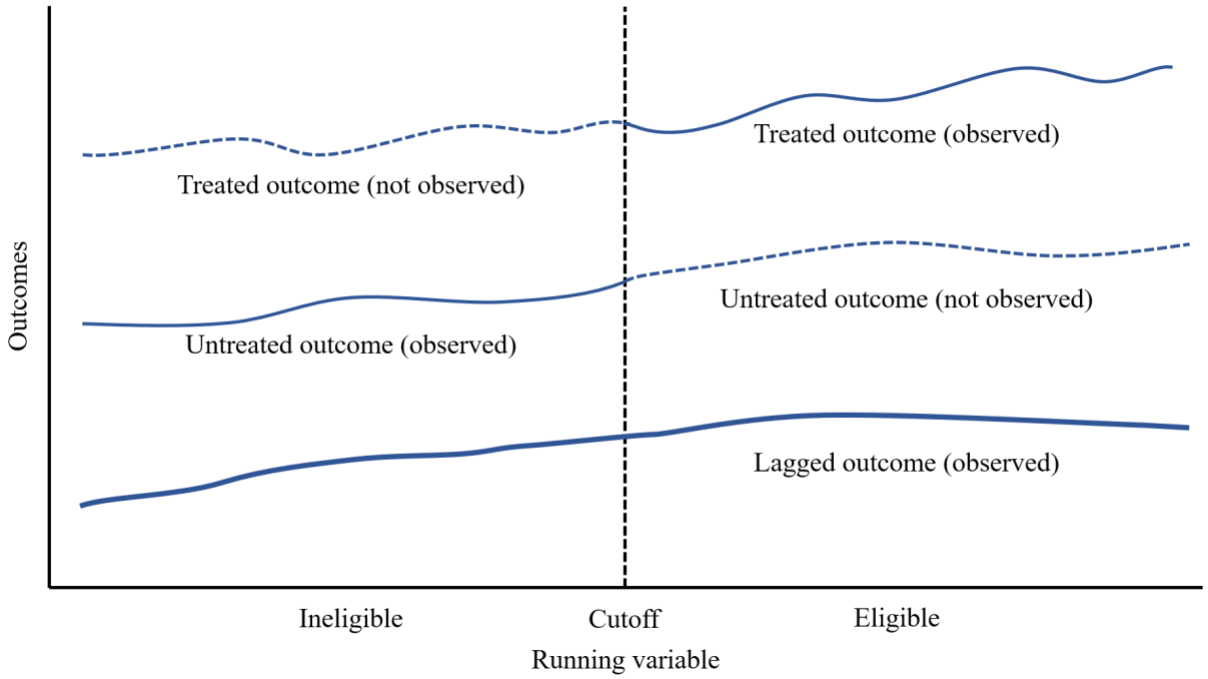
As noted above, we produce a total of 3,600 bias estimates. Most bias estimate (3,200) are without covariates. This is the product of having 2 methods (CRD or RD), 4 estimands, 4 cut points, 5 outcomes, 5 populations, 2 sides of the cut point, and 2 samples (4-150 months and 15-150 months). We produce another 400 bias estimates for CRD using covariates, with cut points at four or five years, and estimands of four years or all years, while the number of outcomes,

populations, sides of the cut point, and samples remains the same. The results with covariates cover fewer bias estimates due to limitations of our aggregate data, as discussed below.

Synthetic Data Generation Method

As noted above, we use synthetic treatment and control groups to estimate our RD and CRD models. In these models, treatment status is determined based on whether or not a cell has a value of the running variable that is above or below a specific cut point. When an RD or CRD model is used outside of a within-study comparison, one only observes treated outcomes on the treated side of the cut point and untreated outcomes on the untreated side. In contrast, when synthetic data are used, we observe the treated and untreated outcomes for all values of the running variable which enables us to estimate the RD, CRD, and RCT models within a single dataset. Figure 1 illustrates a hypothetical dataset. The top line represents outcomes for the treatment group. In this scenario, to create an RD or CRD dataset we drop treated observations below (to the left of) the cut point as indicated by the fact that the line for the treatment group outcome below the cut point is dashed. We also drop untreated outcomes above the cut point. Thus, the second line, which represents outcomes for the control group, is dashed above the cut point. These dashed lines capture potential outcomes in the Rubin causal framework when estimating an RD or CRD model (Rubin 1974). The third line represents lagged outcomes. It combines lagged outcomes for the treatment group above the cut point and the control group below, hence that line is solid on both sides.

Figure 1. *Outcome Measures in Typical RD and CRD Designs*



Data and Samples

We use data from the evaluation of BOND that is particularly well-suited for this analysis. Random assignment was at the individual level and the sample sizes are large, giving us a high level of statistical power. The data include 77,101 treatment observations and 891,429 control observations. In comparison, sample sizes in the Chaplin et al (2018) within-study-comparison of RD ranged from around 100 to 20,000 per study. We use the duration of DI benefit receipt (which we also refer to as "DI duration") at the time of random assignment, measured in months, as the running variable for RD and CRD.[9] We limit our analyses to BOND subjects with 150 months or less of DI duration because of small sample constraints for beneficiaries with longer durations. This covers 75 percent of the sample used in the BOND study.[10] All BOND subjects were enrolled for a minimum of 4 months, hence, DI duration ranges from 4 months to 150 months in some of our analytic samples. We also observe large non-linearities in the relationships between the lagged outcomes and running variable around 15

[9] DI duration is defined as months from DI start to June 1, 2011. See Appendix A for details.
[10] The sample from 4 to 150 months consists of 660,402 control records and 64,426 treatments records. The sample from 15 to 150 months consists of 588,053 control records and 49,842 treatment records. No records are left out of our aggregate data within these ranges. We did not use data above 150 months.

months (see Appendix E). For this reason, we also estimate models using only subjects enrolled for at least 15 months and most of our analyses are based on those results. This covers 66 percent of the sample used in the BOND study.

We focus on SSA administrative data that was used in the evaluation of BOND. BOND included two stages and we focus on Stage 1. The Stage 1 sample is a nationally representative random sample of disabled DI beneficiaries ages 18 to 60 residing in any of 10 SSA Area Office catchment areas as of May 2011. The demonstration randomly assigned more than 77,000 beneficiaries to a Stage 1 treatment group, subject to BOND rules, and nearly 900,000 beneficiaries to a Stage 1 control group, subject to the rules in place at that time. More information on BOND Stage 1 is available in Hoffman et al. (2017) and information on both stages is available in Gubits et al. (2018).

The administrative data include both the Master Earnings File (MEF), which we use to construct employment outcomes, and the Master Beneficiary Record (MBR), which provides information on DI benefit outcomes. We examine outcomes in 2014, the year used as part of a cost-benefit calculation of the program and to predict the size of impacts needed to achieve benefit neutrality (Gubits et al. 2018). The MBR also provides beneficiary information that we use as potential control variables. We measure these control variables either at the start of BOND (in May 2011) or in 2010 for earnings-related measures, because the MEF includes annual, rather than monthly, earnings information. A final source of SSA administrative data is the Supplemental Security Record, which we use for one potential control variable.[11]

In order to protect confidentiality and estimate our models efficiently our analyses are based on aggregated data.[12] We used aggregate data by cell where cells were defined by values of the running variable, treatment status, and a random variable that split each of the units defined by running variable and treatment status in half. For breakdowns by gender and age, this would be a random subset within the gender/age/running variable/treatment status group. For each cell we used the means of the outcomes and covariates, their standard deviations, the sample sizes, and the sums of sample weights used in the BOND evaluation. The random

[11] The database used to administer the Supplemental Security Income program.

[12] The models were estimated at Mathematica. SSA considers individual-level data and small cells that contain fewer than 3 observations as sensitive; such data are only accessible by SSA staff or through other data use agreements. Many of the cells defined by a DI duration of more than 150 months had fewer than 3 observations, hence, we excluded those with more than 150 months duration from analysis. Using aggregate data also enabled us to run the models far more quickly than would be the case if we were to have used individual-level data.

variable split was used so that we could include more covariates in the models.[13] We use those aggregate data to estimate our models following methods recommended by Schochet (2020).

We have confirmed that we can replicate the results of the original BOND study reasonably well using our aggregate data. All differences in impacts between the aggregated analysis and the original BOND study are smaller than 0.01 standard deviations and statistically insignificant. Results are shown in Table 1.

Table 1. *Estimated Impacts in 2014 by Outcome and Source*

Model	Earnings ($)	Employment (%)	Earned BOND yearly amount (%)	DI benefits ($)	DI months (#)
BOND study	16 (28)	0.29* (0.13)	0.20** (0.07)	166*** (26)	0.20*** (0.02)
Aggregate data with covariates	51** (24)	0.49*** (0.14)	0.27*** (0.07)	183*** (32)	0.19*** (0.02)
Aggregate data without covariates	53 (32)	0.51*** (0.20)	0.27*** (0.08)	179*** (65)	0.19*** (0.02)
Standard deviation	6,180	33	16	6,688	3.89

Notes: First line from Gubits et al (2018a). Remaining lines based on analyses by authors. All outcomes are based on annual data for 2014 and based on 2010 covariates. Covariates used with aggregate data are described in Appendix B. */**/*** Impact estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Methods

CRD, RD, and RCT impacts are estimated using methods similar to those of Tang et al (2017). Details are provided in Appendix C. These methods are based on the idea that one can estimate impacts of treatment for the treated group by comparing treated outcomes with predicted counterfactual outcomes for the treatment group (the potential outcomes they would have had if they were not treated). For the RCT we predict counterfactual outcomes using data on the control group; for RD we use the synthetic comparison group outcomes; and for CRD we use the synthetic comparison group outcomes used in RD as well as the lagged outcomes for

[13] The number of cells determines the number of degrees of freedom when running our aggregate models with covariates. Since there are few months of data in many of our models, we can only include a few covariates, as discussed below. We did not create more cells because that would have made it difficult to comply with SSA data access requirements.

both the comparison group and treatment group. To implement these methods, we regress the treated outcomes and the counterfactual estimation outcomes (the outcomes used to predict the treatment group counterfactual) on the running variable and other covariates (if any) in separate regressions for the treatment outcomes and the counterfactual estimation outcomes, and estimate impacts by calculating the differences between the predicted treatment and counterfactual outcomes, predicted for the treatment group. In the CRD and RCT models the treatment group is the full target population. RD models limit the target population to be within the RD bandwidth. All outcomes are standardized to facilitate comparisons across outcomes and so that results can be interpreted in effect size units.

The RD models we estimate are not state of the art. First, we do not control for clustering based on the values of the running variable even though this is often done when estimating RD models (Lee and Card 2008). We did not do this for two reasons. First, doing so would have made it difficult to estimate the standard errors of the bias that we report below. Second, there are concerns about the accuracy of this method when the running variable is discrete (measured in months in our case) since in at least some cases a single unit on the running variable may be larger than the bandwidth needed to obtain reasonably unbiased results (Kolesár and Rothe 2018). This does mean that our estimates of bias for RD may be higher than what they would have been had we controlled for clustering when estimating the standard errors.

The second reason our RD models are not state of the art is that we fix the RD bandwidth at 0.5 years using a model similar to the one used by Tang et al (2017). That is, we include only observations that are within six months from the cut point used to estimate the relationship between the outcome and the running variable. We estimate this model, rather than a more advanced RD model, such as the one proposed by Calonico et al (2020), due to constraints related to our use of aggregate data. More precisely, six months is the smallest bandwidth we felt we could use and still have a reasonable number of degrees of freedom when using the aggregate data.[14] This means that our RD results should not be taken as providing strong evidence regarding the bias that would be obtained using RD compared to RCT estimates of the same estimands. However, because RD estimands may be of less policy interest than estimates that apply to broader populations this is not necessarily a major limitation. Finally, because of

[14] When using aggregate data the degrees of freedom are based on the number of cells used to aggregate the data rather than the original sample size (Schochet 2020).

our use of aggregate data, we are unable to include covariates in our RD models. Recent evidence suggests that correctly estimating RD models with covariates can be complicated even when one has sufficient data (Calonico et al. 2019).

We estimate average bias by group (with the groups described above) by taking averages of our bias estimates within each group. While sampling error influences these results it appears to have a modest impact perhaps because of the large sample sizes of individuals used for each bias estimate and large numbers of bias estimates per group. Almost all groups have at least 99 bias estimates. The only exception is for some of the CRD specification tests. In those cases, we end up with only about 4 bias estimates per group for two groups. For those results the estimates of mean bias could be affected far more by sampling error than for the other groups so we recommend caution when interpreting those results. Almost all of the remaining estimates of average bias are below 0.02 standard deviations.

To estimate the variation in bias within each group of bias estimates we use standard meta-analysis methods used by Weidmann and Miratrix (2021) that adjust for sampling error. Details are provided in Appendix C. As shown in that Appendix we are effectively estimating the variation in bias, $V(B)$, using the variation in observed bias estimates, $v(b)$, minus the variation in observed bias estimates due to sampling, $vs(b)$. We estimate $vs(b)$ using the average of the squared standard errors of the impact estimates, adjusting for the fact that the mean of bias is estimated. These estimates are all weighted by the inverse of the squared standard errors, as is common in meta-analyses. We label our estimate of $V(B)$ as $v(B)$. Thus,

$$(1) \qquad v(B) = v(b) - vs(b)$$

To make the results more interpretable we calculate the square roots of the predicted values from equation 1 and report those in our results in the main body of the report. Those correspond to estimated standard deviations of the bias.

By comparing bias estimates for the groups described above we address the following research questions:

Does the level of estimated bias or variation in bias differ by:

- Model (CRD vs. RD, Table 2),
- Target population (estimand, Table 3),
- Cut point (Table 4),
- Outcome (Table 5),

- Sub-group (Table 6),
- Side of cut point (Table 7),
- CRD specification test outcome (Table 8),
- Covariate inclusion for CRD models (Table 9), or
- Data sample (4-150 months vs. 15-150 months, Table 10).

What are the implications of these results for model selection after one takes into account the precision of the estimates?

Findings

First, we describe average bias and the standard deviation of bias by group (research question 1). We focus on the sample without covariates and with the 15-150 month duration of DI. We then compare results with and without covariates and compare our 15-150 month results with those based on the 4-150 month duration of the DI sample. The tables for each comparison include average bias and the standard deviation of bias. The statistical significance for each standard deviation relative to 0 is in the column labeled "Std Dev" and the statistical significance levels of differences in standard deviations across groups of bias estimates are in the cells under "Statistical significance of Std Dev differences." We conclude with a section on how bias interacts with precision (research question 2).

Bias Results without Covariates

We find that both RD and CRD have a fairly small amount of estimated average bias in results without covariates, but the estimated standard deviations of bias are somewhat higher (Table 2). These results aggregate across 1,600 bias estimates that vary by estimand, cut point, outcome, and side of the cut point. Collectively, estimated average bias (compared against RCT) is below 0.002 in absolute value for both RD and CRD. The estimated standard deviation of bias is below 0.06 for RD and below 0.03 for CRD. The RD versus CRD difference for the estimated standard deviation in bias is statistically significant. For context, in the original BOND RCT, the estimated impacts on earned BOND yearly amount, DI benefits, and months of DI were 0.013, 0.025, and 0.051 respectively in standard deviation units (Gubits et al. 2018). Estimated impacts on the other two outcomes (earnings and employment) were smaller and not statistically significant. Thus, the estimated standard deviation of bias for RD is larger than the estimated

impacts that were statistically significant for 3 outcomes in the original BOND RCT. For CRD that holds true for two outcomes.

Table 2. *Bias for RD vs CRD*

	Bias vs RCT		Statistical significance of Std Dev differences	
Model	**Average**	**Std Dev**	**RD**	**CRD**
RD	-0.0003	0.058***		***
CRD	-0.0019**	0.027***	***	

Note: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (800 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Std Dev means standard deviation. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal represent comparing an estimate to itself. The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

When we look at results grouped based on the estimand (0.5 years from the cut point, two years, four years, all years) we find that estimated average bias remains below 0.01 in absolute value for each estimand (Table 3). The estimated standard deviation in bias is 0.049 for RD with a 0.5 years bandwidth and is smaller, at 0.030, for CRD with a 0.5 years bandwidth. This suggests that CRD actually performs better than the basic RD model when estimating impacts for the 0.5 years bandwidth. As might be expected, RD does somewhat worse as the bandwidth gets larger. In contrast, the estimated standard deviation in bias for CRD actually drops as the bandwidth gets larger.

Table 3. *Bias by Estimand for RD and CRD*

				Statistical significance of Std Dev differences							
		Bias vs RCT		RD				CRD			
Model	Estimand	Average	Std Dev	0.5 years	2 years	4 years	ALL years	0.5 years	2 years	4 years	ALL years
RD	0.5 years	0.0027	0.049***		***	***	***	***			
	2 years	0.0012	0.056***	***		***			***		
	4 years	-0.0028	0.059***	***	***		***			***	
	All years	-0.0048	0.057***	***		***					***
CRD	0.5 years	0.0061***	0.030***	***					***	**	***
	2 years	-0.0037***	0.025***		***			***			**
	4 years	-0.0050***	0.026***			***		**			***
	All years	-0.0003	0.023***				***	***	**	***	

Note: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD 0.5 to CRD at 2 years). The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

The fact that we estimate that the RD models are biased even when using bandwidths of 0.5 years for the RCT might seem surprising since external validity bias is less of an issue there. However, RD models can produce biased estimates even in the absence of external validity bias. This is because, while RD models do produce consistent estimates of impacts at the cut point, they do not produce unbiased estimates of those impacts given finite sample sizes that result in bandwidths that can be substantial in size. This is especially true when the running variable takes on a discrete number of values, as is the case here (Kolesár and Rothe 2018). Thus, even if one focuses on internal validity bias, RD models are likely to be biased at least to some degree. On theoretical grounds, it is quite possible that the results would differ for a more rigorous RD model, for example that of Cattaneo et al (2020), and with a continuous running variable that varied by day, rather than the discrete one we used, that varies only by month. That might be

possible in the context of DI since both disability date and adjudication date vary by day and those events could affect behaviors (Appendix A).

We did not expect to see a great deal of variation in results by cut point, but we do see some (Table 4). Again, there is not much bias on average—the largest estimated average bias by cut point being less than 0.011 standard deviations in magnitude. The estimated standard deviation in bias is somewhat larger for the larger cut points for RD but is lowest for the third cut point for CRD.

Table 4. *Bias by Cut Point for RD and CRD*

				Statistical significance of Std Dev differences							
		Bias vs RCT		RD				CRD			
Model	Cut point in years	Average	Std Dev	2	3	4	5	2	3	4	5
RD	2	0.0026	0.050***			**	***	***			
	3	0.0045	0.049***			***	***		***		
	4	-0.0106	0.064***	**	***					***	
	5	0.0018	0.065***	***	***						***
CRD	2	-0.0055***	0.026***	***						***	
	3	-0.0057***	0.025***		***					***	
	4	0.0002	0.021***			***		***	***		***
	5	0.0017**	0.025***				***			***	

Notes: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD at two years to CRD at four years). The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

When we look at results by outcome, the absolute value of the estimated average bias gets a bit higher, going as high as 0.022 for the fourth outcome for RD, and remains below 0.01

regardless of the outcome for CRD (Table 5). The estimated standard deviation of bias does not vary substantially across outcomes for RD but does vary a bit across outcomes for CRD.

Table 5. *Bias by Outcome for RD and CRD*

				Statistical significance of Std Dev differences									
		Bias vs RCT		RD					CRD				
Model	Outcome	Average	Std Dev	O1	O2	O3	O4	O5	O1	O2	O3	O4	O5
RD	Earnings (O1)	-0.0162***	0.053***						***				
	Employment (O2)	0.0029	0.051***							***			
	Earned BOND yearly amount (O3)	-0.0177***	0.055***								***		
	Amount of benefits (O4)	0.0221***	0.055***									***	
	Months of benefits (O5)	0.0065*	0.053***										***
CRD	Earnings (O1)	0.0010	0.021***	***						***	***	***	**
	Employment (O2)	-0.0065***	0.028***		***				***		***		
	Earned BOND yearly amount (O3)	-0.0039***	0.023***			***			***	***		***	
	Amount of benefits (O4)	0.0022**	0.030***				***		***		***		**
	Months of benefits (O5)	-0.0034***	0.025***					***	**			**	

Notes: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD employment to CRD benefits). The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

When we look at results by demographic group (overall and by gender and age), we see no clear patterns (Table 6). Absolute estimated average bias remains below 0.02 for RD and below 0.01 for CRD. The point estimates for the estimated standard deviation in bias for RD are lower for the full population, at 0.041 than they are for the subgroups, which have estimated

standard deviations ranging from 0.051 to 0.083. As was found for RD, the smallest standard deviation for CRD is for the full population estimates. All of the estimated standard deviations for CRD are below 0.04.

Table 6. *Bias by Group for RD and CRD*

		Bias vs RCT		Statistical significance of Std Dev differences									
				RD					CRD				
Model	Group	Average	Std Dev	ALL	G1	G2	G3	G4	ALL	G1	G2	G3	G4
RD	All	0.0020	0.041***		***	**	***	***	***				
	Young females (G1)	0.0152**	0.083***	***		***		**		***			
	Older females (G2)	-0.0072	0.051***	**	***		***	*			***		
	Young males (G3)	0.0051	0.074***	***		***						***	
	Older males (G4)	-0.0109	0.062***	***	**	*							***
CRD	All	-0.0026***	0.020***	***						***	***	***	**
	Young females (G1)	-0.0006	0.027***		***				***			***	**
	Older females (G2)	-0.0021	0.027***			***			***			***	***
	Young males (G3)	-0.0011	0.039***				***		***	***	***		***
	Older males (G4)	-0.0015	0.022***					***	**	**	***	***	

Notes: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). Outcomes are in standard deviation units. G1 to G4 refer to the 4 subgroups- younger women, older women, younger men, and older men. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD G1 to CRD G4). The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

We see no clear differences in estimated average bias by side of the cut point with all estimates remaining below 0.01 in magnitude (Table 7). The estimated standard deviation in bias is somewhat lower below the cut point than above for both RD and CRD but the difference for RD is not statistically significant.

Table 7. *Bias by Side of Cut point for RD and CRD*

Model	Side of cut point	Bias vs RCT		Statistical significance of Std Dev differences			
				RD		CRD	
		Average	Std Dev	Below	Above	Below	Above
RD	Below	-0.0013	0.056***			***	
	Above	0.0017	0.062***				***
CRD	Below	-0.0037***	0.025***	***			***
	Above	0.0016	0.029***		***	***	

Notes: Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (400 bias estimates per row). Outcomes are in standard deviation units. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD below to CRD above). The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

We find no evidence that using specification tests for CRD would reduce bias within the set of bias estimates we analyzed. More precisely, the tests based on slope differences (ST2 and ST3) almost never fail, and the test of discontinuities (ST1) that does fail fairly often does not predict variation in bias (Table 8). The discontinuity test (ST1) fails just over 10 percent of the time, but failure is not associated with a substantial change in estimated average bias or with a statistically significant change in the estimated standard deviation of bias. The test for whether the slopes of the outcome and lagged outcome are the same only fails for a few cases (ST2). The specification test for whether the slopes for the lagged outcome above and below the cut point are the same fails in only a handful of cases (ST3). Finally, the test that combines all three tests does not predict average bias or variation in bias (ST5). One interpretation of these results is that the discontinuity test (ST1) is not a useful way to test for bias in CRD models and that the results for the non-linearity tests are somewhat ambiguous given that we found so few failures in these data. Our later comparison of results based on the models by numbers of months of DI benefit receipt (Table 10) may be more informative for thinking about the potential importance of non-linearities.

Table 8. *Bias by Specification Test for CRD*

		Bias vs RCT			Statistical significance of Std Dev differences	
Model	**Specification test value**	**Average**	**Std Dev**	**Percentage**	**Fail**	**Pass**
ST1	Fail	-0.0052**	0.026***	12.4		
ST1	Pass	-0.0015**	0.025***	87.6		
ST2	Fail	0.0014	0.061**	0.5		
ST2	Pass	-0.0019**	0.026***	99.5		
ST3	Fail	-0.0687**	0.000	<0.1		
ST3	Pass	-0.0019***	0.026***	>99.9		
ST4	Fail	-0.0015	0.059**	0.5		*
ST4	Pass	-0.0019**	0.027***	99.5	*	
ST5	Fail	-0.0052***	0.025***	12.9		
ST5	Pass	-0.0015	0.027***	87.1		

Notes: Based on 800 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). Outcomes are in standard deviation units. ST1 refers to the test for a discontinuity for the lagged outcome at the cut point. ST2 refers to whether the slope of the outcome is the same as the slope for the lagged outcome on the untreated side of the cut point. ST3 refers to whether the slope of the lagged outcome is the same on both sides of the cut point. ST4 combines tests 2 and 3. ST5 combines tests 1, 2, and 3. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. When the standard deviation is reported as 0.000, this indicates that the observed variation in bias estimates is smaller than what would be expected due to sampling error. Standard errors are provided in Appendix D. Percentage reports the percent of the bias estimates that failed or passed the test. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Bias Results with Covariates

We tested the importance of covariates for a subset of the CRD models. We were not able to test the validity of covariate adjustment for the RD models due to a lack of cells in our aggregated data. We had two random groupings within each value of the running variable. This gives us 12 observations in each regression used to estimate the RD models (one regression for each side of the cut point). We felt that this was not sufficient to test the efficacy of adding covariates to the RD model.

The lack of cells in our aggregate data also constrained our ability to estimate CRD models with covariates. We limited our investigation of the benefits of covariates to CRD models with at least 68 observations on each side of the cut point and used only 4 covariates, in addition to the running variable and the intercept. This gave us over 11 observations per parameter being estimated and 62 degrees of freedom in each of those regressions. This means we only included covariates in CRD models with cut points at four or five years and with bandwidths of four years or all years. Since our data start at 15 months of DI duration, this gives us at least 34 months of data below the four-year cut point. With two cells per month that yields 68 observations (units of aggregated data).

We selected four variables to include as covariates for our regressions: 1) person is a disabled adult child (DAC) beneficiary, 2) person is a dually entitled DAC beneficiary, 3) person is receiving SSI, and 4) person has a legal guardian who was not a representative payee. These four covariates were selected based on the fact that they had the largest absolute t-statistics in regressions of the outcomes on the larger set of covariates we considered and the running variable. These all had average absolute t-statistics above 1.9. The next highest covariate (age squared) had an average absolute t-statistic below 1.5. See Appendix B for details.

Our results suggest no clear benefit to adding covariates to the model (Table 9). Indeed, the absolute estimated average bias and the estimated standard deviation in bias both increase slightly when we add covariates to the model. This may be due in large part to the relatively linear relationships found between the outcomes, lagged outcomes, and running variable for the sample used in Table 9. That is the sample with 15 to 150 months of DI benefits at the time of random assignment, as shown in Appendix E.

Table 9. *Bias by Whether Covariates Are Used for CRD*

	Bias vs RCT		Statistical significance of Std Dev differences	
Model	**Average**	**Std Dev**	**No covariates**	**Covariates**
No covariates	0.0001	0.024***		***
Covariates	-0.0035***	0.026***	***	

Notes: Based on 400 bias estimates from BOND data using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Checking for Importance of Non-Linearities Observed at 15 months

As shown in Appendix E, the relationships between the lagged outcomes, post-tests, and running variable are generally linear above 15 months of DI benefit receipt at the start of the study, at least when viewed across many months. Consequently, we implemented a linear specification for the running variable and limited our analyses to the data from 15 to 150 months. Tang et al (2017) also used a linear specification based on a visual inspection of their data.

While the relationships of the outcomes and lagged outcomes with the running variable are linear above 15 months, the slope from 4 months to 15 months is very different from that above 15 months for all of the lagged outcomes. These non-linearities seems plausible given how our variables were defined. In particular, we would expect a strong positive association between months of DI receipt in 2010 (one of our lagged outcomes) and total months of DI receipt by July of 2011 (our running variable) since the former is a component of the latter for anyone with more than 6 months of DI receipt. At the same time, once one has more than 18 months of DI receipt in July of 2011, additional months may not be correlated with months in 2010 since presumably almost all of those individuals have 12 months in 2010. Hence, we see a drastic change in the slope around that time. A similar pattern might be expected for DI benefits in 2010 since it is highly correlated with months of DI and that is what we find. Similarly, we would expect the opposite for the earnings and employment variables since individuals with higher earnings and employment are less likely to be on DI. Again, this is what we find.

In order to avoid complications associated with these non-linearities we prioritized the models using data from 15 to 150 months. However, we estimated an additional set of models using data from 4 to 150 months and including interactions between the running variable and a dummy variable identifying whether the observation is below the 15 month cut point. Those interactions were designed to allow for the non-linearities we found in the data. As shown in Table 10, the change in sample made very little difference for the RD estimates. However, for CRD the standard deviation in bias rose dramatically, from around 0.027 in the 15-150 month sample, to 0.159 for the sample that goes down to 4 months.[15] This suggests that our attempts to control for the nonlinearities in the sample inclusive of 4 months of DI duration were not sufficient to keep the bias at a low level. This does not mean that it would be impossible to obtain reduced bias for CRD in the presence of such non-linearities, but does suggest that it might be quite challenging.

Table 10. *Bias by Sample for RD and CRD*

				Statistical significance of Std Dev differences			
		Bias vs RCT		4-150 months		15-150 months	
Months of data in sample	**Model**	**Average**	**Std Dev**	**RD**	**CRD**	**RD**	**CRD**
4-150	RD	-0.0009	0.059***		***		
	CRD	-0.1179***	0.159***	***			***
15-150	RD	-0.0003	0.058***				***
	CRD	-0.0019*	0.027***		***	***	

Notes: Based on 3,200 bias estimates from BOND data (800 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Statistical significance results compare the estimated variation in bias between the rows and columns specified. Statistical significance is reported for all unshaded cells. The estimated standard deviation in bias equals the square root of the estimated variation in bias after subtracting variation due to sampling. Standard errors are provided in Appendix D. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

[15] The tests for bias (ST1-ST5) did not do a good job identifying bias in this sample either. The non-linearity tests (ST2 and ST3) performed somewhat better than in the 15-150 sample. More precisely average bias was much larger in the models that failed those tests than in the ones that passed. However, the standard deviation in bias remained large (above 0.15) in both the models that passed and those that failed those tests, as well as in the models that passed and failed the discontinuity tests. Also, the percent passing the non-linearity tests (ST2 and ST3) remained very high, above 96 percent, suggesting that they may not be identifying enough non-linearities to work well.

To summarize, absolute estimated average bias is generally low while the estimated standard deviation in bias is moderate in size when we look at results across the 52 groups of bias estimates we analyzed for the 15 to 150 month sample without covariates. In 50 of those groups average absolute estimated average bias is less than 0.02. In contrast, the estimated standard deviation of bias is above 0.02 in 50 groups and is above 0.04 in 23 groups.

Accounting for Precision of Estimates

When deciding which estimator to use, it is important to consider all potential sources of error including average bias, the standard deviation in bias, and the standard errors of the estimators. Since we found very little evidence of average bias we focus on the standard deviation in bias and the standard errors. To summarize these two quantities, we use their sum:

(2) $v(b)_g = v(B)_g + vs(b)_g$

This is the sum of the estimated variation in bias and the averaged squared standard errors for a set of bias estimates. To make these numbers more interpretable we take the square root and refer to this as the bias-adjusted standard error (BASE).

(3) $BASE(b)_g = \sqrt{v(b)_g}$

Since the standard errors of RD, CRD, and RCT estimators differ, they can have large impacts on the magnitudes of the bias-adjusted standard errors and consequently be an important factor when choosing between methods. We simulated standard errors for RD assuming individual-level data, the same probability of treatment as in BOND, and a uniform distribution around the cut point using estimates based on Schochet (2009).[16] For CRD we assumed that the standard errors were similar to what would be found for an RCT.[17] When we combined the bias estimates with the standard errors, following the $BASE(b)_g$ formula and using the BOND sample sizes, the bias-adjusted standard errors are about the same as the standard deviations of bias because the standard errors, using the full BOND sample, were negligible. Because RCTs have no bias, the bias-adjusted standard errors for RCTs equal their unadjusted standard errors.

[16] Schochet shows that under these assumptions one would need 4 times as many observations as would be needed for an RCT to achieve the same level of statistical power. Using our model and data, we get a ratio of 4.4 on average, across bias estimates, which suggests that an equivalent bias-adjusted standard error could be obtained with an even smaller RCT than what we estimate below.

[17] Using our model and data, the CRD standard errors were about 7 percent larger than the RCT standard errors, on average, which again suggests that an equivalent bias-adjusted standard error could be obtained with a smaller RCT than what we estimate below.

In order to compare the costs of RD and CRD relative to the RCT method we try to estimate how large the treatment group sample size for an RCT would need to be to obtain bias-adjusted standard errors similar to those of an RD or CRD. We focus on the sizes of the treatment groups based on the simplifying assumption that cost is driven only by the size of the treatment group which will often be approximately the case when the treatment itself is expensive relative to business as usual and the comparison group is getting business as usual.[18] Thus, we focus on varying the size of the treatment group. We use the standard deviations of bias for RD and CRD to backout an equivalent RCT that would achieve the same standard errors (and hence same bias-adjusted standard error) using the same sample size for the control group as in the BOND study, but reducing the size of the treatment group. We find that we would need a very large RD (n=77,000 treatment observations) to achieve the same bias-adjusted standard error as a small RCT (n=300 treatment observations). This implies that the RD would be far more expensive than a comparable RCT. Similarly, we would need a very large CRD (again n=77,000 treatment observations) to achieve the same bias-adjusted standard error as a much smaller RCT (n= 1,400 treatment observations). Thus, the RD and CRD studies need very large treatment group sample sizes to achieve bias-adjusted standard errors similar to what could be obtained with much smaller, and thus less expensive, RCTs.

Conclusion

In this study we evaluate the efficacy of regression discontinuity (RD) and comparative RD (CRD) relative to a randomized control trial (RCT), using data from the BOND evaluation. We find little evidence of bias on average but substantial evidence of variation in bias, even after adjusting for sampling variation. We estimate a standard deviation of bias of 0.027 for the CRD estimates. This implies a minimum detectable effect of about 0.075 standard deviations of the outcome, even with a very large sample size, if one uses a bias adjusted standard error (one that incorporates the variation due to bias). This minimum detectable effect is much larger than the impact estimates found in the BOND study, some of which were statistically significant and

[18] In other words, we are assuming that the treatment group costs the government more money. That was not the expectation with BOND but that is what turned out to be the case.

substantively important. This suggests that policymakers should consider using bias-adjusted standard errors when interpreting CRD and RD results.[19]

We find that CRD requires a much larger sample size to achieve the same bias-adjusted standard error as an RCT. This implies that in many cases CRD may not be optimal. That said, there are situations where an RCT may not be feasible. In those situations, CRD may be optimal especially if a standard deviation of bias of 0.027 or larger is acceptable.

For RD we estimate a standard deviation of bias that is even larger than for CRD, at 0.058, which implies a minimum detectable effect of 0.162 standard deviations of the outcome. This estimate of the standard deviation of bias incorporates the fact that RD is only estimating treatment effects at the cut point, while policy makers are generally interested in impacts for larger populations with impacts that may vary by the values of the running variable. However, even when the RD and RCT use the same bandwidth, the standard deviation in bias is still substantial at 0.049 suggesting that RD models similar to the one we estimated may be less useful in practice than CRD. That said, it is not always possible to obtain the data needed for a CRD model. In our case we added a lagged outcome to the standard RD model. Other variables could be used in place of the lagged outcome but if none are available than an RD model may still be optimal.

Our results appear to be reasonably consistent with prior literature. In particular, Weidmann and Miratrix (2021) estimated a standard deviation of bias of 0.04 for models estimated using a simple matching algorithm.[20] This is fairly similar to our estimate of about 0.03 for CRD. Chaplin et al. (2018) estimated a standard deviation of bias for RD models of 0.07. In comparison, we estimate about 0.06 for RD. This also suggests that outside of the context of DI, our results may be more encouraging for both CRD and RD models as these levels of bias were judged to be more acceptable in the prior literature cited here.

Our estimates of bias come with several caveats. Bias might be larger than what we estimate because WSCs can only capture some forms of bias. For example, manipulation of the running variable is not possible in a WSC like the one we used. In addition, researcher bias for or against an intervention or towards finding statistically significant impact estimates is far less likely in a WSC than in a standard RD or CRD study. In the opposite direction, bias for RD

[19] Similar ideas have been proposed by Ganong and Jager (2018) and by Deke, Finucane, and Thal (2022).
[20] This is based on communication with the authors. In the paper they reported the mean of absolute bias.

models may be lower than what we estimate. This is because we estimated relatively simplistic RD models. More sophisticated models may yield far less bias.

While CRD models appear to have promise, attention should be paid to non-linearities between the running variable and the outcome or lagged outcome. In this analysis, we observed large changes in the slopes of the lagged outcomes on the running variable around 15 months of DI receipt. There are plausible explanations for these relationships. However, they could generate significant bias in CRD models that use data below 15 months of DI receipt, which is what we found when we included those data in our models. We also found evidence suggesting that standard specification tests might do little to alleviate this source of bias.

It is important to note that these conclusions are based on RD and CRD conducted with a discrete running variable and the BOND data; hence, results may not generalize to continuous running variables, other types of data, or other periods in time. Nevertheless, we think they provide a valuable first step to developing a better understanding of the trade-offs between RD, CRD, and RCTs in situations similar to those considered here.

More research on variations of the CRD model would enhance our understanding of the performance of the approach relative to RCTs. For example, there are methods similar to CRD that match the comparison group on one side of the cut point to the treatment group on the other side (Angrist and Rokkanen 2015). Such matching may improve on the simple regression adjustment used here. In addition, the running variable we used was not continuous: the start of DI receipt always occurs on the first day of a given month. Other measures of duration of DI receipt, such as the adjudication date of a DI application, may occur on any day of the month and might yield smaller standard errors for the CRD models and less bias.

It may also be helpful to do more research on the RD model. RD performed better than CRD when there were non-linearities in the relationship between the running variable and the lagged outcome. Hence, it might be helpful to assess the performance of RD models using methods proposed by Calonico et al (2020) and using a running variable that varied by day, rather than month, in order to obtain more precise estimates. It might also be worth exploring how bias varies with the sample sizes used as that could have important impacts on bias especially for RD.

References

Angrist, Joshua D. and Miikka Rokkanen. 2015. "Wanna Get Away? Regression Discontinuity Estimation of Exam School Effects Away from the Cutpoint." *Journal of the American Statistical Association* 110(512): 1331-1344.

Bell, Stephen H., Daniel Gubits, David Stapleton, David Wittenburg, Michelle Derr, Arkadipta Ghosh, and Sara Ansell. 2011. "BOND Implementation and Evaluation: Evaluation Analysis Plan." Final report submitted to Social Security Administration. Cambridge, MA: Abt Associates. Available at: https://www.ssa.gov/disabilityresearch/documents/BOND%20Evaluation%20Analysis%20Plan.pdf

Calonico, Sebastian., Matias D. Cattaneo, and Max H. Farrell. 2020. "Optimal Bandwidth Choice for Robust Bias Corrected Inference in Regression Discontinuity Designs." *The Econometrics Journal* 23: 192-210.

Calonico, Sebastian, Matias D. Cattaneo, Max H. Farrell, and Rocío Titiunik. 2019. "Regression Discontinuity Designs using Covariates." *The Review of Economics and Statistics* 101(3): 442-451.

Cattaneo, M. D., B. Frandsen, and R. Titiunik. 2015. "Randomization Inference in the Regression Discontinuity Design: An Application to Party Advantages in the U.S. Senate," *Journal of Causal Inference* 3: 1-24.

Chaplin, Duncan D., Thomas D. Cook, Jelena Zurovac, Jared S. Coopersmith, Mariel M. Finucane, Lauren N. Vollmer, and Rebecca E. Morris. 2018. "The Internal and External Validity of the Regression Discontinuity Design: A Meta-Analysis of 15 Within-Study Comparisons." *Journal of Policy Analysis and Management* 37(2): 403-429.

Chen, Susan, and Wilbert van der Klaauw. 2008. "The Work Disincentive Effects of the Disability Insurance Program in the 1990s." *Journal of Econometrics* 142(2): 757-784.

Cook, Thomas D., Shadish, William R., and Wong, Vivian C. "Three Conditions under which Experiments and Observational Studies Produce Comparable Causal Estimates: New Findings from Within-study Somparisons." *Journal of Policy Analysis and Management* 27: 724-750.

Deke, J., M. Finucane, and D. Thal. 2022. "The BASIE (BAyeSian Interpretation of Estimates) Framework for Interpreting Findings from Impact Evaluations: A Practical Guide for Education Researchers." NCEE 2022-005. Washington, DC: U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance.

Deshpande, Manasi. 2016. "Does Welfare Inhibit Success? The Long-Term Effects of Removing Low-Income Youth from the Disability Rolls." *American Economic Review* 106(11): 3300-3330.

Ganong, Peter and Simon Jäger. 2018. "A Permutation Test for the Regression Kink Design." *Journal of the American Statistical Association* 113 (522): 494-504.

Gelber, Alexander, Timothy J. Moore, and Alexander Strand. 2017. "The Effect of Disability Insurance Payments on Beneficiaries' Earnings." *American Economic Journal: Economic Policy* 9(3): 229-261.

Gelber, Alexander, Timothy Moore, and Alexander Strand. 2015. "The Effect of Disability Insurance on Beneficiaries' Mortality." Working Paper No. NB 14-06. Cambridge, MA: National Bureau of Economic Research.

Gleason, Phil, Alex Resch, and Jillian Berk. 2018. "RD or Not RD: Using Experimental Studies to Assess the Performance of the Regression Discontinuity Approach." *Evaluation Review* 41(1): 3-33.

Gubits, Daniel, David Stapleton, Stephen Bell, Michelle Wood, Denise Hoffman, Sarah Croake, David R. Mann, Judy Geyer, David Greenberg, Austin Nichols, Andrew McGuirk, Meg Carroll, Ustav Kattel, and David Judkins. 2018. "BOND Implementation and Evaluation: Final Evaluation Report." Submitted to Social Security Administration. Cambridge, MA: Abt Associates. Available at: https://www.ssa.gov/disabilityresearch/documents/BOND%20Deliv%2024e2%20FER%20Vol%201%2020181018.pdf

Hoffman, Denise, Sarah Croake, David R. Mann, David Stapleton, Priyanka Anand, Chris Jones, Judy Geyer, Danny Gubits, Stephen Bell, Andrew McGuirk, David Wittenburg, Debra Wright, Amang Sukasih, David Judkins, and Michael Sinclair. 2017. "BOND Implementation and Evaluation: 2016 Stage 1 Interim Process, Participation, and Impact Report." Submitted to Social Security Administration. Cambridge, MA: Abt Associates. Available at: https://www.ssa.gov/disabilityresearch/documents/BOND_Deliverable%2024c%202%201_3%2020%2017_clean_toSSA_with_Warning.pdf

Kisbu-Sakarya, Yasemin, Thomas D. Cook, Yang Tang, and M.H. Clark. 2018. "Comparative Regression Discontinuity: A Stress Test with Small Samples." *Evaluation Review* 42(1): 111-143.

Kolesár, Michal and Christoph Rothe. 2018. "Inference in Regression Discontinuity Designs with a Discrete Running Variable," *American Economic Review* 108(8): 2277-2304.

LaLonde, Robert J. 1986. "Evaluating the Econometric Evaluations of Training Programs with Experimental Data." *American Economic Review* 76(4): 604-620.

Lee, David S. and David Card. "Regression Discontinuity Inference with Specification Error." *Journal of Econometrics* 142: 655-674.

Liu, Su and David C. Stapleton. 2011. "Longitudinal Statistics on Work Activity and Use of Employment Supports for New Social Security Disability Insurance Beneficiaries." *Social Security Bulletin* 71(3): 35-60.

McCann, Clare. 2019. "Closing the Evidence Gap: Doing More of What Works in Higher Education." Washington, DC: New America. Available at: https://www.newamerica.org/education-policy/reports/closing-evidence-gap/introduction/

Olsen, Robert B., Stephen H. Bell and Austin Nichols. 2018. "Using Preferred Applicant Random Assignment (PARA) to Reduce Randomization Bias in Randomized Trials of Discretionary Programs." *Journal of Policy Analysis and Management* 37(1): 167-180.

Rubin, Donald B. 1974. "Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies." *Journal of Educational Psychology* 66(5): 688-701.

Schochet, Peter Z. 2020. "Analyzing Grouped Administrative Data for RCTs Using Design-Based Methods." *Journal of Educational and Behavioral Statistics* 45(1): 32-57.

Schochet, Peter Z. 2009. "Statistical Power for Regression Discontinuity Designs in Education Evaluations." *Journal of Educational and Behavioral Statistics* 34: 238-266.

Stapleton, David C., Stephen H. Bell, David C. Wittenburg, Brian Sokol, and Debi McInnis. 2010. "BOND Implementation and Evaluation: BOND Final Design Report." Final report submitted to the Social Security Administration, Office of Program Development and Research. Cambridge, MA: Abt Associates. Available at: https://www.ssa.gov/disabilityresearch/documents/BOND_Design%20Report_FINAL_Del2-2_12-17-10.pdf

Stuart, Elizabeth A., Stephen H. Bell, Cyrus Ebnesajjad, Robert B. Olsen, and Larry L. Orr. 2017. "Characteristics of School Districts that Participate in Rigorous National Educational Evaluations." *Journal of Research on Educational Effectiveness* 10(1): 168-206.

Tang, Yang, Thomas D. Cook, and Yasemin Kisbu-Sakarya. 2018. "Statistical Power for the Comparative Regression Discontinuity Design with a Non-Equivalent Comparison Group." *Psychological Methods* 23(1): 150-168.

Tang, Yang, Thomas D. Cook, Yasemin Kisbu-Sakarya, Heinrich Hock, and Hanley Chiang. 2017. "The Comparative Regression Discontinuity (CRD) Design: An Overview and Demonstration of Its Performance Relative to Basic RD and the Randomized Experiment." *Advances in Econometrics* 38: 237-279.

Tipton, E., Jessaca Spybrook, Kaitlyn G. Fitzgerald, Qian Wang, and Caryn Davidson. 2021. "Toward a System of Evidence for All: Current Practices and Future Opportunities in 37 Randomized Trials." *Educational Researcher* 50(3): 145-156.

U.S. Social Security Administration. 2019. "Annual Statistical Report on the Social Security Disability Insurance Program, 2018." Publication No. 13-11826. Washington, DC. Available at: https://www.ssa.gov/policy/docs/statcomps/di_asr/2018/di_asr18.pdf

Weidmann, Ben and Luke Miratrix. 2021. "Lurking Inferential Monsters? Quantifying Selection Bias in Evaluations of School Programs." *Journal of Policy Analysis and Management* 40: 964-986.

Wing, Coady and Thomas D. Cook. 2013. "Strengthening the Regression Discontinuity Design Using Additional Design Elements: A Within-Study Comparison." *Journal of Policy Analysis and Management* 32(4): 853-877.

Wong Vivian C. and Peter M. Steiner. 2018. "Designs of Empirical Evaluations of Nonexperimental Methods in Field Settings." *Evaluation Review* 42(2): 176-213.

Appendix A. Running Variable

In this Appendix, we provide more details on the running variable used in the within study comparison: the duration of DI benefit receipt at the time of BOND enrollment. We use the same measure of benefit duration as used in the implementation and evaluation of Stage 1 of BOND. Specifically, this is the number of days from the approximate start of DI benefits to June 1, 2011, rounded to the nearest month. The DI start date is based on the award date for 93 percent of the sample that has an award date on file, and the entitlement date for the remaining 7 percent of our sample. We describe the DI award date and entitlement date, as well as other milestones on the path to benefit receipt to better understand these dates and related alternatives (Table A1). We distinguish between continuous variables that vary by day and discrete ones, that can only take on certain values, usually the first day of each month.

Table A1. *Social Security Disability Insurance Dates*

Date	Continuous?	Definition
Disability onset date	Yes	The first date in which a person meets the medical and non-medical criteria for DI.
Entitlement date	No	Date of initial entitlement to DI benefits; also considered to be the enrollment date. Occurs 5 months after the first day of the month following the disability onset date.
Adjudication date	Yes	Date SSA made an initial adjudication or appeal decision on a claim. SSA notifies beneficiaries of their decision, typically through a mailed letter. This occurs after the disability onset date, but could be before or after the entitlement date.
Award date	No	Date a beneficiary is entitled to benefit payments and received their first payment. Payments are typically made on the third of the month or the second, third or fourth Wednesday of the month.

The measure of DI duration used in the evaluation of BOND and used as the running variable in this analysis is based on discrete variables (award date and entitlement date). However, RDs are likely to have less bias and both RDs and CRDs are likely to have smaller standard errors when based on continuous variables. The implication of using a discrete versus a continuous variable is that CRD and RD might perform better relative to an RCT. Alternatives

to consider are disability onset date, which is a continuous variable in the MBR, or adjudication date, which is a continuous variable recorded in SSA's Data Analysis Support Hub (DASH).

Appendix B. Covariates

The use of aggregate data imposes limits on the number of covariates we can include in our models and which models we can run with covariates. In the end, we limit our analyses of covariates to the models with the cut point equal to four or five years and to those with bandwidths of four years or all years. We select four covariates, which allows for 62 degrees of freedom, by running regressions of each of the five outcomes on all of the covariates and the running variable using the original data by cell (15-150 months, not by age or gender), calculating the average t-statistics across the five outcomes, and picking the four covariates with the largest average t-statistics.

We start with the same set of covariates used in the BOND study, except that we omit covariates based on the running variable, lagged outcomes, and closely related variables and interactions. Covariates omitted for these reasons include monthly benefit amount (MBA) at baseline, Average Indexed Monthly Earnings (AIME) as of May 2011, earnings in 2010 (the year prior to BOND random assignment year), if monthly benefit amount (MBA) at baseline is equal to zero, if AIME as of May 2011 are equal to zero, the interaction of monthly benefit amount at baseline and AIME as of May 2011, AIME as of May 2011 (squared), and if any employment in 2010 (the year prior to random assignment year).

Some covariates were omitted because they were either too similar to our running variable or part of the running variable vector, R_p. These include 36 months or fewer of DI receipt (short duration), if DI start date is on or after January 1, 2010 (very short duration), number of years receiving DI, number of years receiving DI squared, interaction of very short-duration x 2010 earnings, and the interaction of age and number of years receiving DI.

The covariates considered for inclusion in X_p are age, age squared, the county 2010 employment rate for people with a disability, the county April 2011 unemployment rate, and dummies for SSI receipt, if a disabled adult child (DAC) beneficiary, if a dually entitled DAC beneficiary, if a disabled widow(er) beneficiary (DWB), if a dually entitled DWB, if missing employment rate for people with a disability, if disabled, if female, if has neoplasms, if has mental disorders, if has back or other musculoskeletal issues, if has nervous system disorders, if has circulatory system disorders, if has genitourinary system disorders, if has a respiratory disorder, if has severe visual impairments if has issues with the digestive system, if has other impairments, if has unknown impairments, if has a representative payee, if lives in a rural area, if

missing the 2010 unemployment rate/rural status, has auxiliary beneficiary (AUX) who is not a DAC or DWB, receives written beneficiary notices in Spanish, ineligible for Stage 2 of the BOND study for geographical reasons, and ineligible for Stage 2 for having a legal guardian who was not a representative payee.

Appendix C. Methods

Datasets

In order to describe our methods more completely we first describe a series of datasets we use in our estimation. These include the original data, a target population dataset, a CRD dataset, an RD dataset, and an RCT dataset. We create the target population dataset first from the original data, and then use that to create the other datasets. This is done to ensure that we standardize the data in the same way across models within a target population. The target population dataset is usually limited to individuals with between 15 and 150 months of DI duration at the start of BOND in most of our analyses, but as discussed earlier, we also estimate bias using individuals with as few as 4 months to see how much including those additional months of data matters. In addition, for most of our bias estimates we further limit the target population based on gender, age, or other criteria.

The RD dataset is designed to simulate data we would have in a real RD situation in which there was no RCT. The target population data include both treatment and control observations from the original data. We create the RD dataset by dropping treatment observations on the untreated side of the cut point and control observations on the treated side from the target population dataset. Thus, the remaining data has only treated observations on the treated side of the cut point and only control observations on the untreated side. These control observations are used as the comparison observations in the RD models.

The CRD dataset augments the RD dataset in a way that facilitates using lagged outcomes in the counterfactual regressions. We do this by creating an additional record for each individual—both treatment and control—in which the dependent variable is the lagged outcome. This enables us to run regressions to predict the counterfactual where the counterfactual estimation outcomes include both the outcomes for the comparison group (on the untreated side of the cut point) as well as lagged outcomes for both the treatment and comparison groups (on both sides of the cut point).

The RCT dataset is similar to the RD dataset but designed to facilitate estimation of impacts using RCT methods. To do this we take the target population dataset and drop all observations on the untreated side of the cut point which leaves us with both treatment and control observations on the treated side of the cut point. Thus, the RCT dataset can be used to estimate treatment on the treated effects.

Standardizing

We standardize our outcomes, lagged outcomes, and covariates for two different purposes. The outcomes and lagged outcomes are standardized so that our results are comparable across outcomes. Our covariates are standardized to facilitate estimating treatment and counterfactual outcomes for the treatment groups.

We standardize all outcomes and lagged outcomes using the corresponding means and standard deviations for the full BOND study control sample. Specifically, we subtract the corresponding mean and divide by the standard deviation.

We standardize the covariates and the running variable for CRD by mean-centering them based on their means in the treatment sample (on the treated side of the cut point). This is done once for each target population using only the data within the relevant distance from the cut point for the model being estimated. We use the same mean-centering when estimating the corresponding RCT model. We standardize the running variable for RD by mean-centering it based on its mean on the treated side of the cut point within the 0.5 year bandwidth. However, when estimating the corresponding RCT model we use a bandwidth appropriate for the target population (0.5 years, two years, four years, or all years). Thus, the RCT is designed to estimate impacts for the full target population while the RD always estimates impacts at the cut point so the difference between the RD and RCT estimates is in part because they are estimating different estimands.

Comparative Regression Discontinuity Method

We estimate impacts using the CRD model by subtracting predicted counterfactual outcomes from predicted treated outcomes. To predict the counterfactuals, we regress the counterfactual estimation outcomes in the CRD data file (the outcomes for the comparison group and the lagged outcomes for both the treatment and comparison groups) on the running variable, the control variables (if any), and an indicator for whether the dependent variable is an outcome rather than a lagged outcome[21]:

$$(C.1) \quad Y_{pr} = \alpha_{uCRD} + \beta_{uxCRD} X_p + \beta_{upre} L_{pr} + e_{upr}$$

where

[21] This is similar to equation 1 in Tang et al (2017).

- Y_{pr} is the dependent variable *r* for person *p* (either an outcome, when *r*=0, or a lagged outcome, when *r*=1),
- X_p is the set of covariates for person *p* (when included in the model) and the running variable, RV_p, and is mean-centered on the values for the relevant treatment group, as discussed earlier,
- L_{pr} is an indicator identifying that the record is for a lagged outcome (1) rather than for an outcome (0), and
- e_{upr} is an error term for an untreated record.

We use this regression to estimate the counterfactual outcome that would be observed for each treatment group member had they not been treated. It is the predicted counterfactual outcome for treatment subjects conditional on their values of the running variable and the covariates when setting the lagged outcome indicator variable (P_{pr}) to 0. Since we mean-center the covariates at the values for the treatment group, the intercept, $\alpha_{uCRD,}$ is the average of the predicted values across treatment group members.[22]

We run a similar regression using the outcomes for the CRD treatment group in the CRD analysis file[23]:

(C.2) $$Y_{pr} = \alpha_{tCRD} + \beta_{txCRD} X_p + e_{tpr}$$

where e_{tpr} is an error term for a treated outcome record rather than a counterfactual one, so *r*=0.

Doing this enables us to estimate the impacts for the treatment group as follows:

(C.3) $$\beta_{CRD} = \alpha_{tCRD} - \alpha_{uCRD}$$

Randomized control trial method

We estimate RCT impacts using equations that are very similar to the ones used to estimate the CRD impacts, except that we use the RCT data in which both the treatment and comparison groups are on the treated side of the cut point and all dependent variables are outcomes, so *r*=0.

[22] To see this, note that the intercept represents the outcome for someone with 0 values for each covariate. Since the covariates are mean-centered for the treatment group that means that the intercept estimates the outcome for someone with the mean values of the covariates. Since we are using a linear regression, the predicted value of the mean is the same as the mean of the predicted value. Thus, the intercept is the mean of the predicted counterfactual for the treatment group.

[23] This is similar to equation 2 in Tang et al. (2017).

(C.4) $Y_{pr} = \alpha_{uRCT} + \beta_{uxRCT}X_p + e_{upr}$

(C.5) $Y_{pr} = \alpha_{tRCT} + \beta_{txRCT}X_p + e_{tpr}$

Again, this equation allows the impact to vary with the values of the covariates and running variable. Doing this enables us to estimate the average impacts for the treatment group as follows:

(C.6) $\beta_{RCT} = \alpha_{tRCT} - \alpha_{uRCT}$

Finally, we can estimate bias as:

(C.7) $b_{CRD} = \beta_{CRD} - \beta_{RCT}$

We run these regressions using the cell-level means weighted by the cell-level sums of weights. When estimating equation C.7 we account for the covariance in estimates between equations C.3 and C.6. Using this method enables us to estimate CRD and RCT impacts, biases, and their standard errors.

Comparative Regression Discontinuity Specification Tests

We conduct specification tests for the CRD counterfactual models by adding coefficients to equation C.1 designed to capture discontinuities in the lagged outcome at the cut point, differences in slopes between the outcome and lagged outcome on the untreated side of the cut point, and differences in slopes across the cut point for the lagged outcome.

(C.8) $Y_{pr} = \alpha_u + \beta_{u1}RV_p + \beta_{u2}X_p + \beta_{u3}O_p + \beta_{u4}TRDD_p + \beta_{u5}TRDD_pRV_p + \beta_{u6}O_pRV_p + e_{pr}$

where $TRDD_p$ is a dummy variable set to 1 for the treated side of the cut point and 0 otherwise, and O_{pr} equals 1 minus L_{pr}. Thus, it identifies if the record is an outcome and not a lagged outcome.

The parameter β_{u4} identifies any discontinuity in the lagged outcome at the cut point. The parameter β_{u5} identifies any difference in slopes for the lagged outcome above and below the cut point and the coefficient β_{u6} identifies any difference in slopes between the outcome and the lagged outcome on the untreated side of the running variable. We use specification tests based on each of these alone as well as on various combinations of them. More precisely we ask:

1. Is there a discontinuity in the lagged outcome at the cut point?

2. Is there a difference between the slopes on the running variable for the untreated outcome and lagged outcome?
3. Is there a difference between the slopes on the running variable for the lagged outcome above and below the cut point?
4. Do either of the slopes differ?
5. Do any of these tests reject the model?

Thus, we are able to identify (1) the efficacy of each of these specification tests on their own, (2) the marginal benefits of the discontinuity test conditional on the slopes test, and (3) the marginal benefits of each slope test conditional on the other. We categorize tests 1, 2, and 3 as having failed if the value is greater than 0.05 standard deviations and is statistically significant at the 0.05 value. Tests 4 and 5 combine results from 1, 2, and 3.

Regression Discontinuity Design Method

We estimate the RD models using methods similar to those used for CRD. As noted earlier, this is a simplistic RD model. Rather than estimating a bandwidth (proximity of the observations to the cut point), we selected 0.5 years.

A major difference between the RD and CRD models is that when we are estimating bias associated with using an RD to estimate impacts outside of the 0.5 years bandwidth (the parts of the population within two years, four years or all years from the cut point) we still use the 0.5 years bandwidth to define the estimand for the RD. This is accomplished by using the RD dataset with the covariates mean centered based on their values for the treatment group in the 0.5 years bandwidth and not based on a larger bandwidth. To clarify this distinction, we create a set of covariates, XRD, that is mean centered by the values of the covariates for the treatment group within the 0.5 years bandwidth of the cut point. This gives us the following equations based only on outcome data (r=0).

(C.9) $Y_{pr} = \alpha_{uRD} + \beta_{uxRD}XRD_p + e_{upr}$

(C.10) $Y_{pr} = \alpha_{tRD} + \beta_{txRD}XRD_p + e_{tpr}$

Using this equation, we can estimate the RD and bias as follows:

(C.11) $\beta_{RD} = \alpha_{tRD} - \alpha_{uRD}$

(C.12) $b_{RD} = \beta_{RD} - \beta_{RCT}$

Methods to Estimate and Analyze Bias

We use standard meta-analytic methods to estimate equation 1 in the main body of our report based on the methods used by Weidmann and Miratrix (2021).

(1) $v(B) = v(b) - vs(b)$

Their paper estimates bias for fourteen treatments with three outcomes per treatment. They adjust for correlation due to sampling across bias estimates for different outcomes within the same treatment and assume the bias estimates are independent across treatments. We allow all of the bias estimates in our original sample to be correlated due to sampling error. In order to obtain sets of independent bias estimates we based our analyses on 20 sets, *s*, of replicates, *r*. Each replicate is a dataset equal in size to the original BOND data, and drawn with replacement from those data at the individual level, with stratification by treatment status and the values of the running variable. Each set of replicates has 5 replicates, so we have a total of 100 replicates across all sets. We allow for correlation of bias estimates within replicate and assume that the bias estimates are independent across replicate. For a given group, *g*, of bias estimates and set of replicates, *s*, the formula for $v(B)$[24] is:

(C.13) $v(B)_{gs} = [Q_{gs}-(K_g-1)]/D_{gs} = Q_{gs}/D_{gs} - (K_g-1)/D_{gs}$

where

- Q_{gs} = $\Sigma(b_{gjs} - Mb_{gs})^2 * W_{gjs}$ across *j*,
- b_{gjs} = bias estimate *j* in group *g* for set *s*,
- W_{gjs} = weight = $1/SE^2_{gjs}$,
- SE^2_{gjs} = squared standard error for b_{gjs},
- Mb_{gs} = mean of b_{gjs} across *j* for group *g* and set *s*, weighted by W_{gjs},
- D_{gs} = denominator = $SW_{gs} - (SW^2_{gs}/SW_{gs})$,
- SW_{gs} = sum of weights = $\Sigma(W_{gjs})$ across *j*,
- SW^2_{gs} = sum of squared weights = $\Sigma(W^2_{gjs})$ across *j*,
- K_g = effective sample size of group = $k_g * w/[1+(k_g-1)*ICC_g]$,
- k_g = average number of bias estimates in group across replicates,[25]

[24] $V(B)$ is often written as t^2 in this literature. In addition, when the sampling error is estimated to be larger than the observed variation in bias estimates the estimate of $V(B)$ is set to 0. We also report an estimate of 0 for the one group where this happens.

[25] For most groups k_g does not vary across replicates. However, for the groups where we test for discontinuities and differences in slopes it does.

- w = 5 = number of replicates in each set,
- ICC_g = $\max(0, (SE^2Mb_g - (MSE^2_g/(k_g*w)))/[MSE^2_g])$,[26]
- SE^2Mb_g = $V(Mb_{gs})$ = variation in Mb_{gs} across sets within the group,[27] and
- MSE^2_g = mean of SE^2_{gjs} for group g, weighted by W_{gjs}.

The first component of equation 13, Q_{gs}/D_{gs}, is the weighted variance of the raw bias estimates (not adjusted for sampling error). The second component, $(K_g-1)/D_{gs}$, is an estimate of sampling error. Thus, the difference gives us an estimate of the variation in bias after subtracting variation due to sampling.

We use our results to analyze how much E(B) and $v(B)$ differ depending on the characteristics of the sample used to create the bias estimates, as discussed below. We select characteristics that can be used to divide the sample and corresponding bias estimates into non-overlapping sets. To analyze how the level of bias differs with these characteristics, we compare the mean bias and variance in bias between each of these non-overlapping groups.

Our key parameter estimates are the estimates of mean bias and variation in bias for different groups of bias estimates, and differences in the estimated variation in bias between certain groups of bias estimates. We use bootstrapping to estimate the standard errors of these parameter estimates. More precisely, we estimate mean bias, the variation in bias, and differences in those quantities between the non-overlapping groups using for each of the 20 sets of replicates. The estimated standard error of each parameter estimate of interest is equal to the standard deviation of the estimates of that parameter across those bootstrap samples.

[26] Note that SE_2Mb_g is not included in the denominator because MSE^2_g captures total variation by itself, including the variation between replicates. All estimates of the ICCs were positive except when we had no bias estimates.
[27] We cannot estimate this parameter by set since it is based on variation across sets. Thus, our final standard errors do not account for variation across sets in this parameter or K_g.

Appendix D. Variation in Bias Estimates

In this Appendix we present the standard errors of the key parameter estimates presented in the main body of the report. Those parameter estimates cover mean bias, variation in bias, and differences in the variation in bias between subgroups.

Table D1. *Bias for RD vs CRD and Standard Errors of Key Parameters*

	Bias vs RCT		Standard errors of Var Bias differences	
Model	**Average**	**Var Bias**	**RD**	**CRD**
RD	-0.0003 (0.0033)	0.0034*** (0.0002)		0.0002
CRD	-0.0019** (0.0009)	0.0007*** (0.0000)	0.0002	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (800 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Std Dev means standard deviation. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal represent comparing an estimate to itself. The estimated variation in bias equals the variation in estimated bias after subtracting variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D2. *Bias by Estimand for RD and CRD and Standard Errors of Key Parameters*

				Standard errors of Var Bias differences							
		Bias vs RCT		RD				CRD			
Model	Estimand	Average	Var Bias	0.5 years	2 years	4 years	All	0.5 years	2 years	4 years	ALL
RD	0.5 years	0.0027 (0.0030)	0.0024*** (0.0001)		0.0002	0.0002	0.0002	0.0001			
	2 years	0.0012 (0.0031)	0.0032*** (0.0002)	0.0002		0.0001	0.0001		0.0002		
	4 years	-0.0028 (0.0036)	0.0034*** (0.0002)	0.0002	0.0001		0.0001			0.0002	
	All	-0.0048 (0.0040)	0.0032*** (0.0002)	0.0002	0.0001	0.0001					0.0002
CRD	0.5 years	0.0061*** (0.0016)	0.0009*** (0.0001)	0.0001					0.0001	0.0001	0.0001
	2 years	-0.0037*** (0.0010)	0.0006*** (0.0000)		0.0002			0.0001		0.0000	0.0000
	4 years	-0.0050*** (0.0010)	0.0007*** (0.0000)			0.0002		0.0001	0.0000		0.0000
	All	-0.0003 (0.0010)	0.0005*** (0.0000)				0.0002	0.0001	0.0000	0.0000	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD 0.5 years to CRD at two years). The estimated variation in bias equals the variation in estimated bias minus the variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D3. *Bias by Cut point for RD and CRD and Standard Errors of Key Parameters*

				Standard errors of Var Bias differences							
		Bias vs RCT		RD				CRD			
Model	Cut point in years	Average	Var Bias	2	3	4	5	2	3	4	5
RD	2	0.0026 (0.0046)	0.0025*** (0.0003)		0.0004	0.0006	0.0006	0.0003			
	3	0.0045 (0.0047)	0.0024*** (0.0003)	0.0004		0.0006	0.0006		0.0003		
	4	-0.0106 (0.0069)	0.0041*** (0.0006)	0.0006	0.0006		0.0007			0.0006	
	5	0.0018 (0.0050)	0.0042*** (0.0005)	0.0006	0.0006	0.0007					0.0005
CRD	2	-0.0055*** (0.0007)	0.0007*** (0.0001)	0.0003					0.0001	0.0000	0.0001
	3	-0.0057*** (0.0009)	0.0006*** (0.0000)		0.0003			0.0001		0.0001	0.0001
	4	0.0002 (0.0008)	0.0004*** (0.0000)			0.0006		0.0000	0.0001		0.0000
	5	0.0017** (0.0008)	0.0006*** (0.0000)				0.0005	0.0001	0.0001	0.0000	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD at 2 years to CRD at 4 years). The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D4. *Bias by Outcome for RD and CRD and Standard Errors of Key Parameters*

				Standard errors of Var Bias differences by outcome									
		Bias vs RCT		RD					CRD				
Model	Outcome	Average	Var Bias	O1	O2	O3	O4	O5	O1	O2	O3	O4	O5
RD	Earnings (O1)	-0.0162*** (0.0049)	0.0028*** (0.0004)		0.0005	0.0003	0.0006	0.0007	0.0004				
	Employment (O2)	0.0029 (0.0045)	0.0026*** (0.0003)	0.0005		0.0004	0.0005	0.0005		0.0003			
	Earned BOND yearly amount (O3)	-0.0177*** (0.0048)	0.0030*** (0.0003)	0.0003	0.0004		0.0005	0.0005			0.0003		
	Amount of benefits (O4)	0.0221*** (0.0037)	0.0030*** (0.0004)	0.0006	0.0005	0.0005		0.0005				0.0004	
	Months of benefits (O5)	0.0065* (0.0038)	0.0028*** (0.0005)	0.0007	0.0005	0.0005	0.0005						0.0005
CRD	Earnings (O1)	0.0010 (0.0013)	0.0004*** (0.0001)	0.0004						0.0001	0.0000	0.0001	0.0001
	Employment (O2)	-0.0065*** (0.0013)	0.0008*** (0.0001)		0.0003				0.0001		0.0001	0.0001	0.0001
	Earned BOND yearly amount (O3)	-0.0039*** (0.0014)	0.0005*** (0.0001)			0.0003			0.0000	0.0001		0.0001	0.0001
	Amount of benefits (O4)	0.0022** (0.0011)	0.0009*** (0.0001)				0.0004		0.0001	0.0001	0.0001		0.0001
	Months of benefits (O5)	-0.0034*** (0.0011)	0.0006*** (0.0001)					0.0005	0.0001	0.0001	0.0001	0.0001	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD employment to CRD benefits). The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D5. *Bias by Group for RD and CRD and Standard Errors of Key Parameters*

				Standard errors of Var Bias differences									
		Bias vs RCT		RD					CRD				
Model	Group	Average	Var Bias	ALL	G1	G2	G3	G4	ALL	G1	G2	G3	G4
RD	All	0.0020 (0.0041)	0.0017*** (0.0002)		0.0010	0.0004	0.0007	0.0007	0.0002				
	Young females (G1)	0.0152** (0.0067)	0.0068*** (0.0010)	0.0010		0.0012	0.0014	0.0013		0.0010			
	Older females (G2)	-0.0072 (0.0050)	0.0026*** (0.0003)	0.0004	0.0012		0.0008	0.0007			0.0003		
	Young males (G3)	0.0051 (0.0065)	0.0055*** (0.0007)	0.0007	0.0014	0.0008		0.0011				0.0008	
	Older males (G4)	-0.0109 (0.0067)	0.0039*** (0.0007)	0.0007	0.0013	0.0007	0.0011						0.0007
CRD	All	-0.0026*** (0.0007)	0.0004*** (0.0000)	0.0002						0.0001	0.0001	0.0002	0.0001
	Young females (G1)	-0.0006 (0.0017)	0.0007*** (0.0001)		0.0010				0.0001		0.0001	0.0002	0.0001
	Older females (G2)	-0.0021 (0.0015)	0.0007*** (0.0000)			0.0003			0.0001	0.0001		0.0002	0.0001
	Young males (G3)	-0.0011 (0.0013)	0.0015*** (0.0002)				0.0008		0.0002	0.0002	0.0002		0.0002
	Older males (G4)	-0.0015 (0.0009)	0.0005*** (0.0001)					0.0007	0.0001	0.0001	0.0001	0.0002	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). Outcomes are in standard deviation units. G1 to G4 refer to the 4 subgroups- younger women, older women, younger men, and older men. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD G1 to CRD G4). The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D6. *Bias by Side of Cut Point for RD and CRD and Standard Errors of Key Parameters*

Model	Side of cut point	Bias vs RCT		Standard errors of Var Bias differences			
				RD		CRD	
		Average	Var Bias	Below	Above	Below	Above
RD	Below	-0.0013 (0.0045)	0.0031*** (0.0004)		0.0005	0.0004	
	Above	0.0017 (0.0043)	0.0038*** (0.0003)	0.0005			0.0003
CRD	Below	-0.0037*** (0.0014)	0.0006*** (0.0000)	0.0004			0.0000
	Above	0.0016 (0.0011)	0.0008*** (0.0001)		0.0003	0.0000	

Notes: Numbers in parentheses are standard errors. Based on 1,600 bias estimates from BOND data without covariates using the 15-150 month sample (400 bias estimates per row). Outcomes are in standard deviation units. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The shaded cells on the diagonal of the cells in the upper left and lower right are cells that represent comparing an estimate to itself. The shaded cells in the upper right and lower left are shaded because they represent comparisons that are not of substantive interest (for example, RD below to CRD above). The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D7. *Bias by Specification Test for CRD and Standard Errors of Key Parameters*

Test	Test outcome	Bias vs RCT			Standard errors of Var Bias differences	
		Average	Var Bias	Percentage	Fail	Pass
ST1	Fail	-0.0052** (0.0021)	0.0007*** (0.0001)	12.4		0.0001
	Pass	-0.0015** (0.0006)	0.0006*** (0.0000)	87.6	0.0001	
ST2	Fail	0.0014 (0.0178)	0.0037** (0.0018)	0.5		0.0019
	Pass	-0.0019** (0.0008)	0.0007*** (0.0000)	99.5	0.0019	
ST3	Fail	-0.0687** (0.0274)	- -	<0.1		.
	Pass	-0.0019*** (0.0007)	0.0007*** (0.0000)	>99.9	.	
ST4	Fail	-0.0015 (0.0189)	0.0035** (0.0015)	0.5		0.0014
	Pass	-0.0019** (0.0009)	0.0007*** (0.0000)	99.5	0.0014	
ST5	Fail	-0.0052*** (0.0015)	0.0006*** (0.0001)	12.9		0.0001
	Pass	-0.0015 (0.0009)	0.0007*** (0.0000)	87.1	0.0001	

Notes: Numbers in parentheses are standard errors. Based on 800 bias estimates from BOND data without covariates using the 15-150 month sample (160 bias estimates per row). Outcomes are in standard deviation units. ST1 refers to the test for a discontinuity for the lagged outcome at the cut point. ST2 refers to whether the slope of the outcome is the same as the slope for the lagged outcome on the untreated side of the cut point. ST3 refers to whether the slope of the lagged outcome is the same on both sides of the cut point. ST4 combines tests 2 and 3. ST5 combines tests 1, 2, and 3. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. A "-" indicates that the estimated standard deviation of bias is 0. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D8. *Bias by Whether Covariates Are Used for CRD and Standard Errors of Key Parameters*

	Bias vs RCT		Standard errors of Var Bias differences	
Model	**Average**	**Var Bias**	**No covariates**	**Covariates**
No covariates	0.0001 (0.0007)	0.0006*** (0.0000)		0.0000
Covariates	-0.0035*** (0.0010)	0.0007*** (0.0000)	0.0000	

Notes: Numbers in parentheses are standard errors. Based on 400 bias estimates from BOND data using the 15-150 month sample (200 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Table D9. *Bias by Sample for RD and CRD and Standard Errors of Key Parameters*

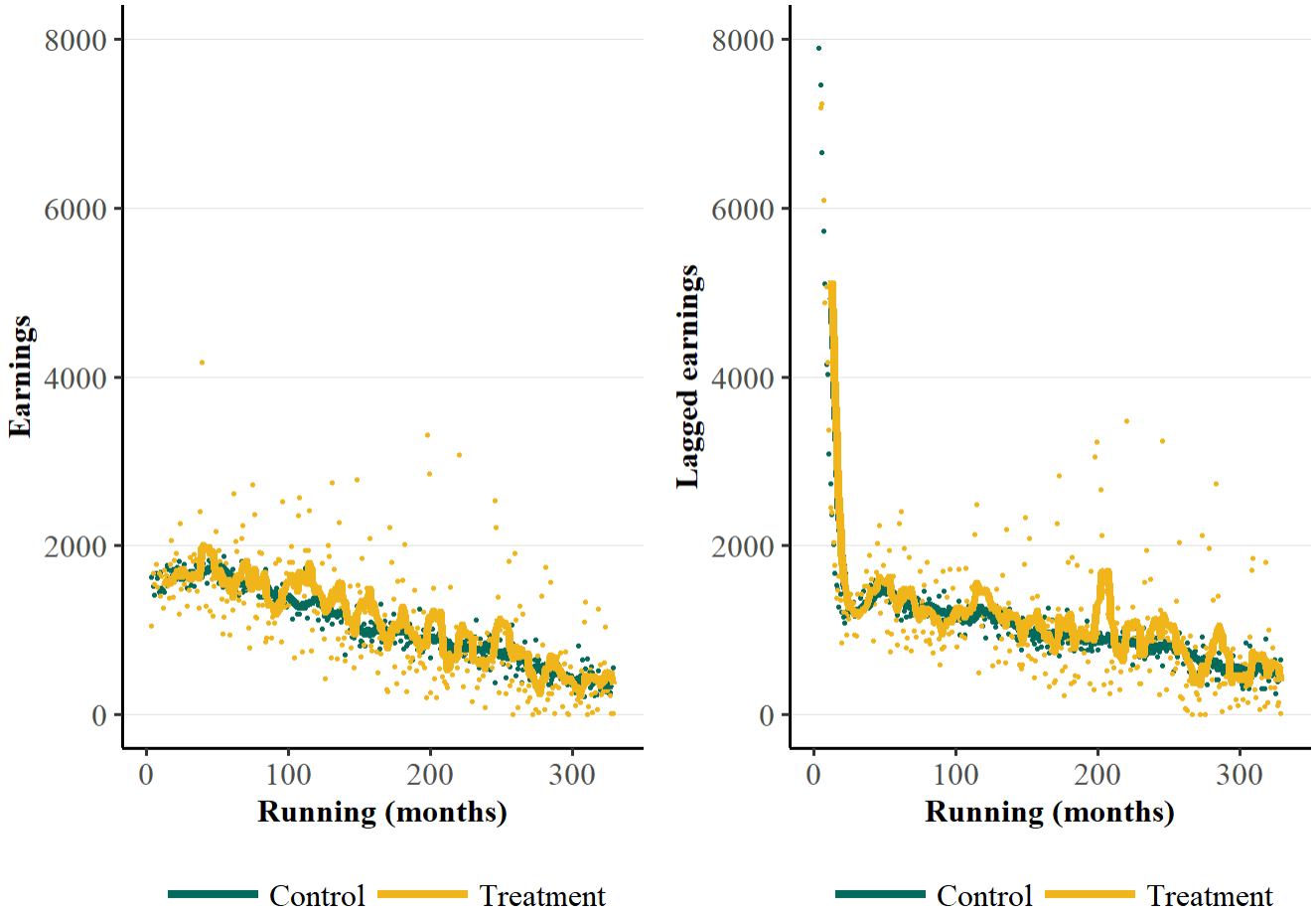
		Bias vs RCT		Standard errors of Var Bias differences			
				4-150 months		15-150 months	
Months included in sample	**Estimand**	**Average**	**Var Bias**	**RD**	**CRD**	**RD**	**CRD**
4-150	RD	-0.0009 (0.0032)	0.0035*** (0.0002)		0.0005	0.0000	
	CRD	-0.1179*** (0.0026)	0.0253*** (0.0004)	0.0005			0.0005
15-150	RD	-0.0003 (0.0032)	0.0034*** (0.0002)	0.0000			0.0002
	CRD	-0.0019* (0.0011)	0.0007*** (0.0000)		0.0005	0.0002	

Notes: Numbers in parentheses are standard errors. Based on 3,200 bias estimates from BOND data (800 bias estimates per row). Outcomes are in standard deviation units. RD means regression discontinuity. CRD means comparative RD. Each bias estimate compares an RD or CRD estimate to a randomized control trial (RCT) estimate. Standard errors of Var Bias differences results compare the estimated variation in bias between the rows and columns specified. The estimated variation in bias equals the variation in estimated bias minus variation due to sampling. The 4-150 sample covers people with values of 4 to 150 months on the running variable (DI eligibility receipt). */**/*** Estimate is significantly different from zero at the .10/.05/.01 levels, respectively, using a two-tailed t-test.

Appendix E. Graphs of Data

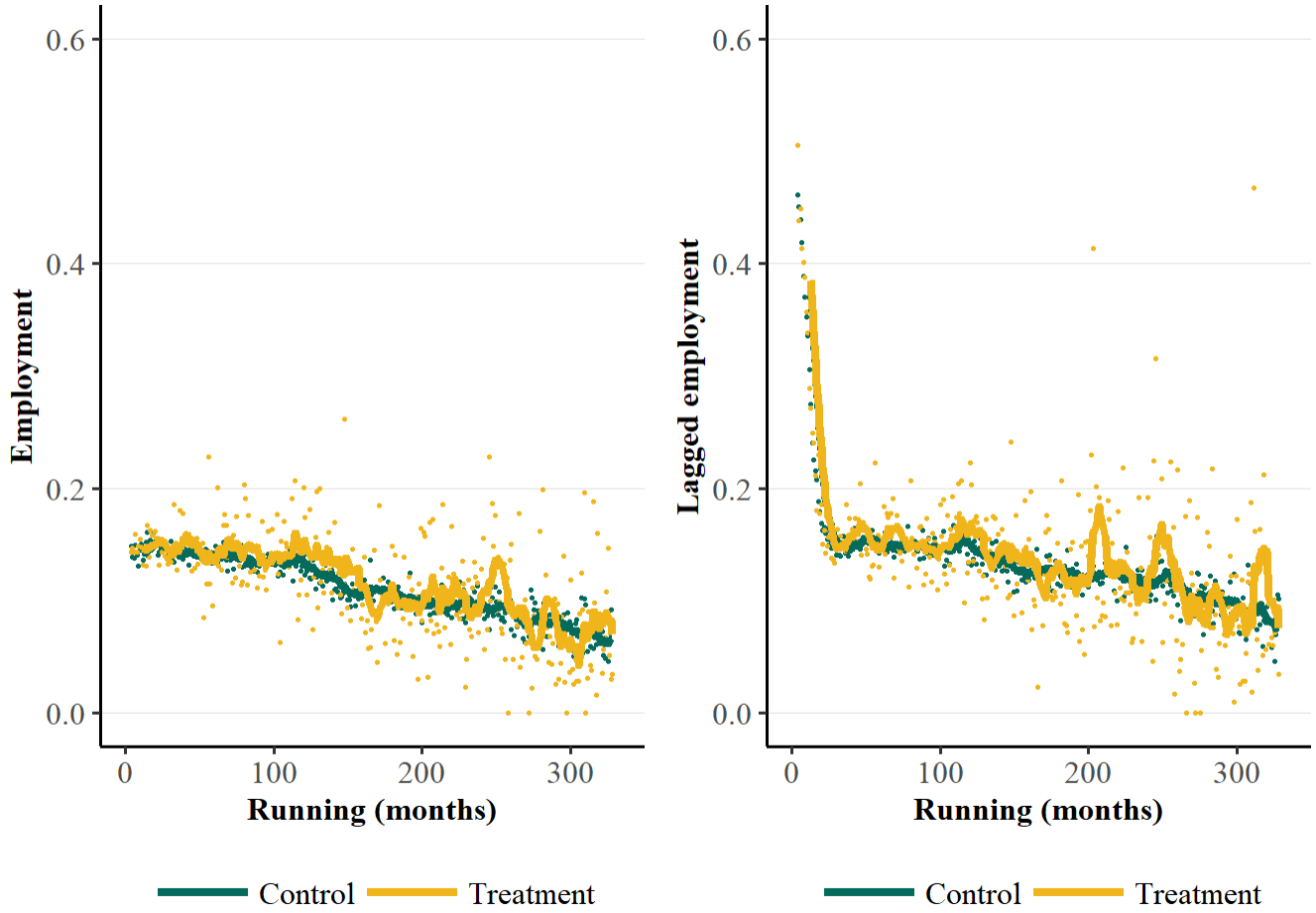
In this Appendix we present figures showing the relationships between the running variable, the outcomes, and the lagged outcomes. The figures show the outcomes and lagged outcomes in separate graphs, by the values of the running variable and treatment status with running averages superimposed over the raw data. These figures help to illustrate how the relationships between the outcomes and running variable are generally fairly linear, the sharp discontinuity in the lagged outcomes around 15 months, and the amount of noise found in the aggregate data, which is more pronounced for the treatment group than the control group. This makes sense because the treatment group is much smaller. An explanation for the sharp discontinuity around 15 months is given in the main body of the report.

Figure E1. *Earnings by Values of the Running Variable*



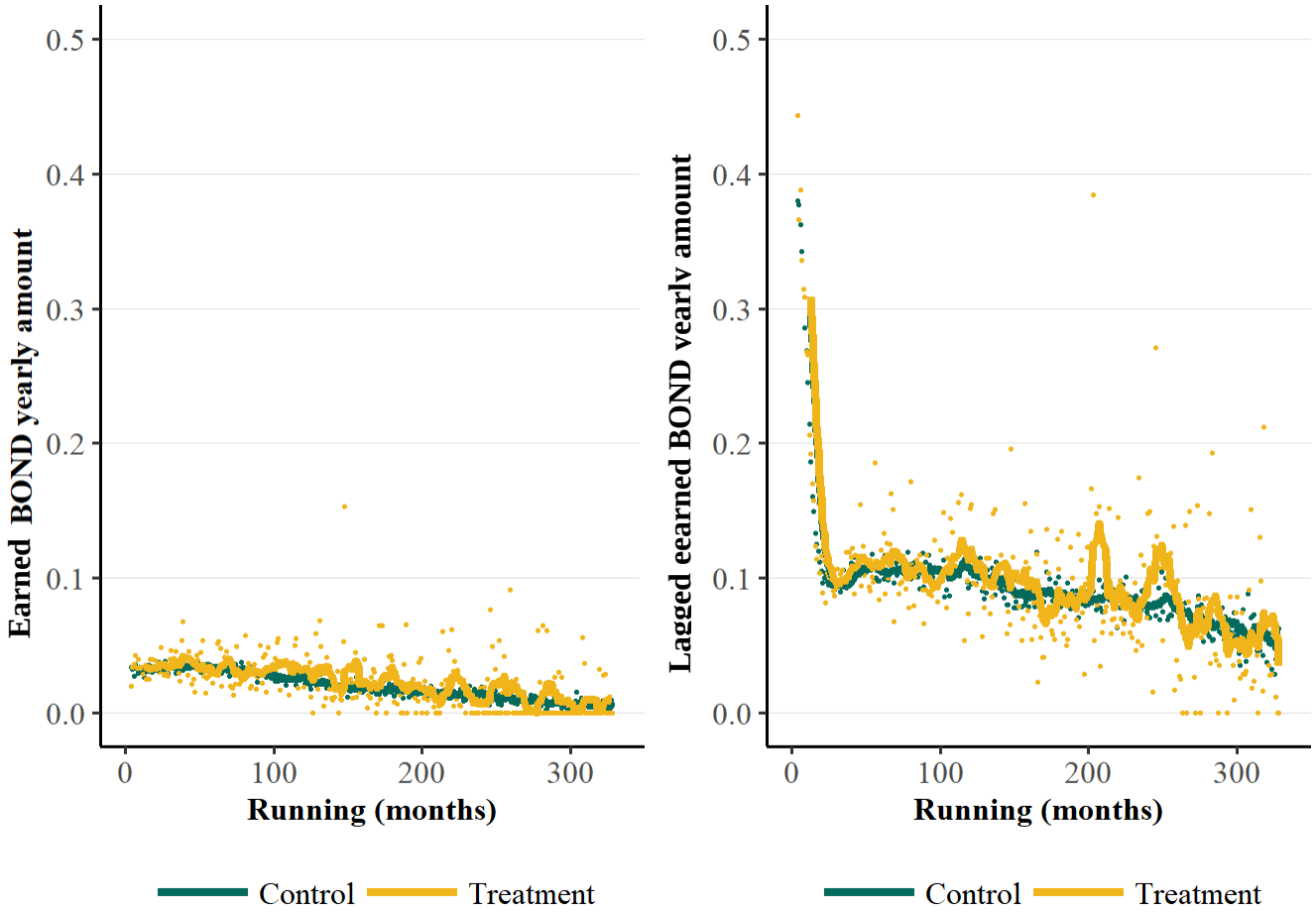
Notes: Outcomes are for 2014. Lagged outcome is for 2010. Running variable is lifetime months of DI as of June 2011.
Source: BOND data.

Figure E2. *Employment by Values of the Running Variable*



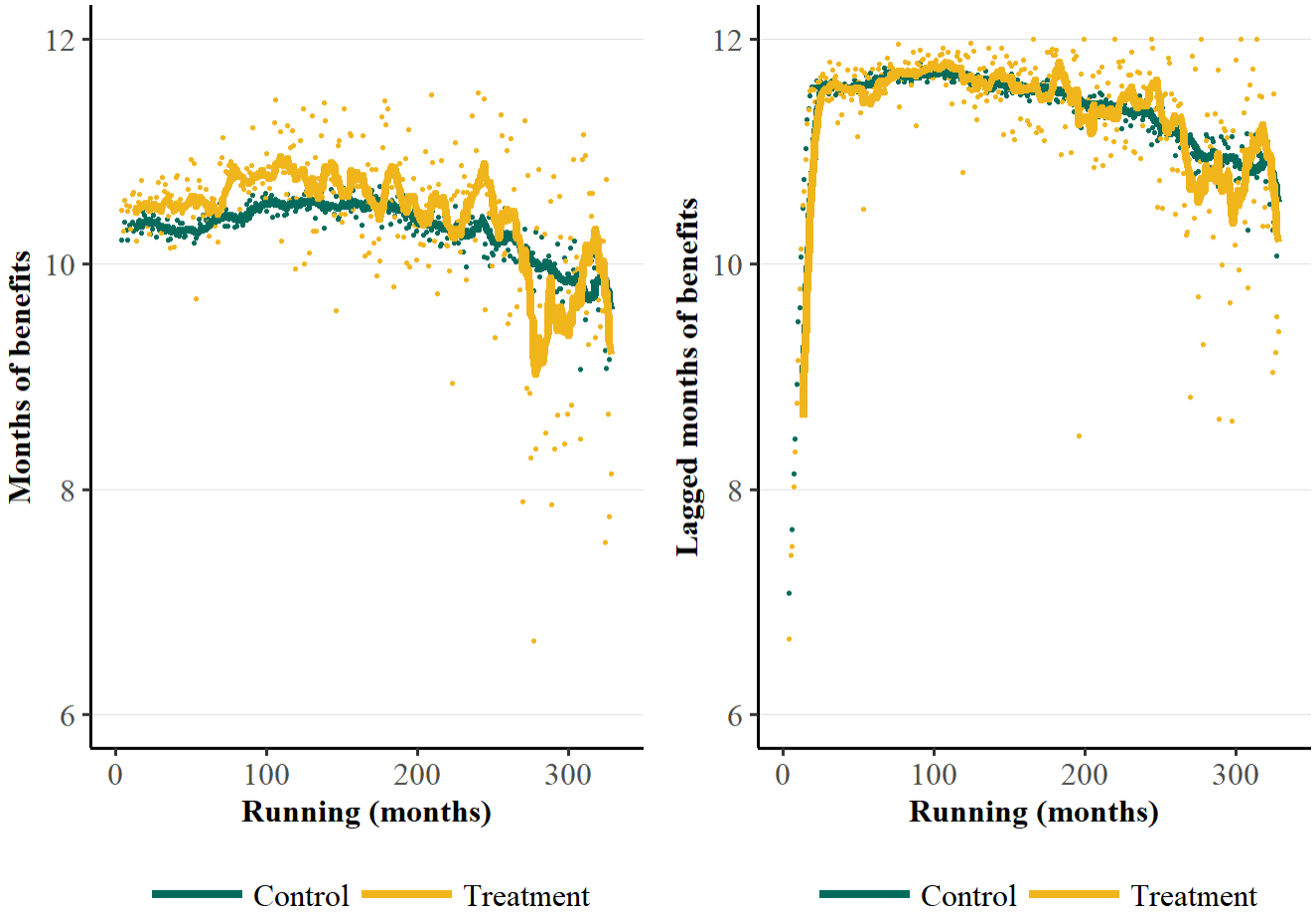
Notes: Outcomes are for 2014. Lagged outcome is for 2010. Running variable is lifetime months of DI as of June 2011.
Source: BOND data.

Figure E3. *Earned BOND Yearly Amount by Values of the Running Variable*



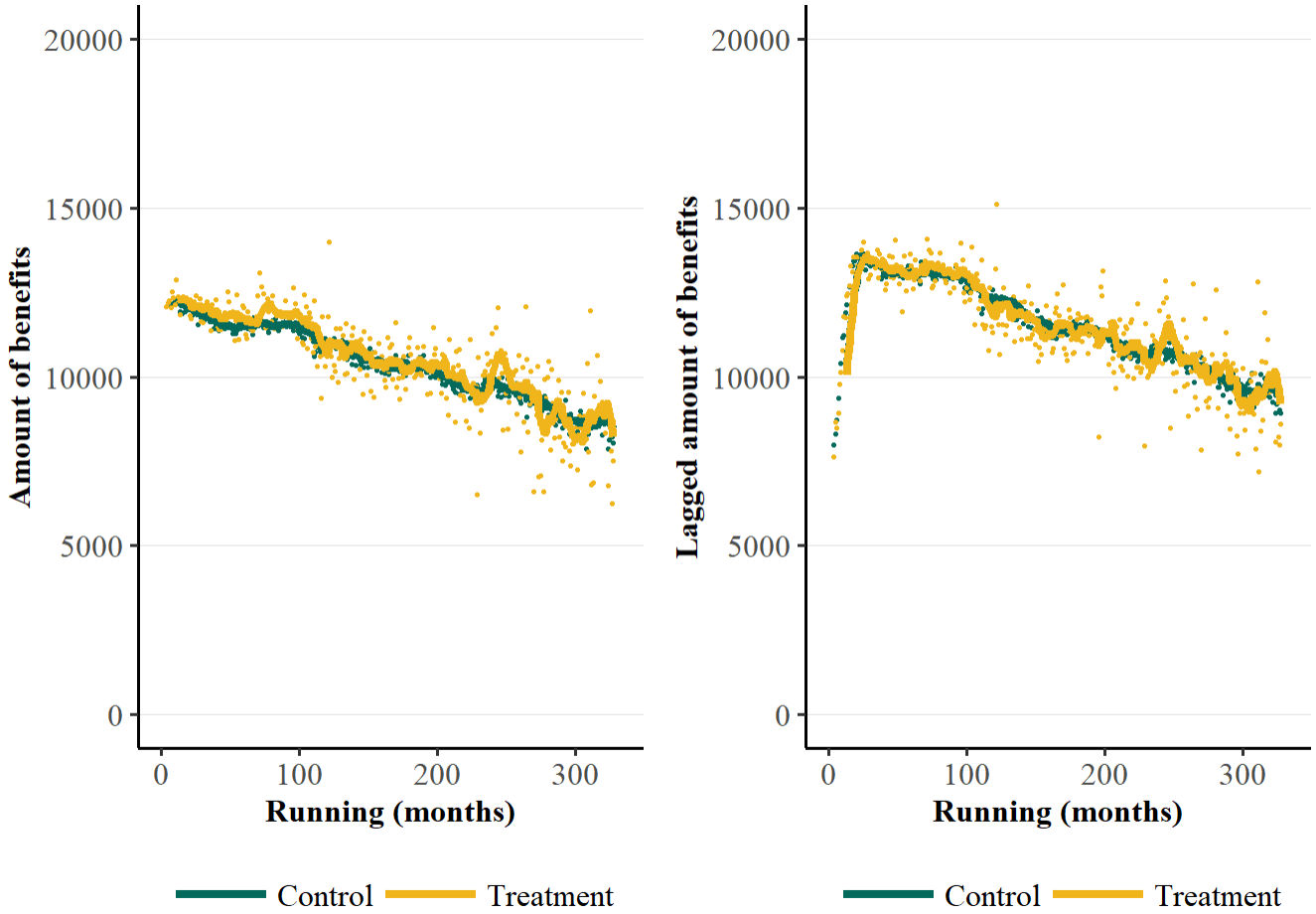
Notes: Outcomes are for 2014. Lagged outcome is for 2010. Running variable is lifetime months of DI as of June 2011.
Source: BOND data.

Figure E4. *DI Months by Values of the Running Variable*



Notes: Outcomes are for 2014. Lagged outcome is for 2010. Running variable is lifetime months of DI as of June 2011.
Source: BOND data.

Figure E5. *DI Benefits by Values of the Running Variable*



Notes: Outcomes are for 2014. Lagged outcome is for 2010. Running variable is lifetime months of DI as of June 2011.
Source: BOND data.

RECENT WORKING PAPERS FROM THE CENTER FOR RETIREMENT RESEARCH AT BOSTON COLLEGE

Work Overpayments Among New Social Security Disability Insurance Beneficiaries
Denise Hoffman, Monica Farid, Serge Lukashanets, Michael T. Anderson, and John T. Jones, July 2022

What Is the Relationship Between Deprivation and Child SSI Participation?
Michael Levere, David Wittenburg, and Jeffrey Hemmeter, May 2022

What Share of Noncovered Public Employees Will Earn Benefits that Fall Short of Social Security?
Jean-Pierre Aubry, Siyan Liu, Alicia H. Munnell, Laura D. Quinby, and Glenn Springstead, April 2022

Employer Concentration and Labor Force Participation
Anqi Chen, Laura D. Quinby, and Gal Wettstein, March 2022

Will the Jobs of the Future Support an Older Workforce?
Robert L. Siliciano and Gal Wettstein, March 2022

Employment Outcomes for Social Security Disability Insurance Applicants Who Use Opioids
April Yanyuan Wu, Denise Hoffman, Paul O'Leary, and Dara Lee Luca, February 2022

Would 401(k) Participants Use a Social Security "Bridge" Option?
Alicia H. Munnell and Gal Wettstein, December 2021

The Alignment Between Self-Reported and Administrative Measures of Application to and Receipt of Federal Disability Benefits in the *Health and Retirement Study*
Jody Schimmel Hyde and Amal Harrati, December 2021

Changes in New Disability Awards: Understanding Trends and Looking Ahead
Lindsay Jacobs, December 2021

The Influence of Early-Life Economic Shocks on Aging Outcomes: Evidence from the U.S. Great Depression
Valentina Duque and Lauren L. Schmitz, December 2021

Are There "Hot Spots" of Primary Impairments among New SSDI Awardees – and Do We Know Why?
Jody Schimmel Hyde, Anna Hill, Jonathan Schwabish, and Aaron R. Williams, December 2021

All working papers are available on the Center for Retirement Research website (https://crr.bc.edu) and can be requested by e-mail (crr@bc.edu) or phone (617-552-1762).